Specifying Latent Structure Characteristics in Mixed-membership Models

Ramnath Balasubramanyan - rbalasub@cs.cmu.edu

CMU-LTI-13-007

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave, Pittsburgh, PA 15213
www.lti.cs.cmu.edu

Thesis Committee:
William W. Cohen (Chair), Carnegie Mellon University
Noah A. Smith, Carnegie Mellon University
Stephen Fienberg, Carnegie Mellon University
Padhraic Smyth, University of California, Irvine

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy
In Language and Information Technologies

Acknowledgments

The process of working on this PhD has been challenging on both technical and personal terms. I am deeply grateful for the extensive support and inspiration I have had the privilege of enjoying, from a whole host of people.

I am greatly indebted to my advisor William Cohen, who has patiently guided over the past six years, as I frequently stumbled. While I have leaned on him technically extensively, my gratitude towards him extends far beyond guidance with technical matters. Watching him frame research in a bigger context and finding connections to other areas have been valuable lessons which I'm sure will help me in my role as a researcher.

I am also thankful to the other members on my thesis committee. Noah Smith's energy and passion for research is always an inspiration. I am thankful to Steve Fienberg for thoughtful comments about the statistical validity of the ideas I've worked on. Padhraic Smyth has been very supportive and his comments and advice have greatly helped me shape my thesis.

My wife Beena, has been a rock during this long journey through grad school. I have learnt a lot from her over the years from her about being perseverent and relentless. She is singularly responsible for keeping me on track. I am eternally grateful to my family, especially my parents for teaching me the value of education and forever putting my interests above theirs.

A big part of the grad school experience is interactions with fellow students. These friendships and the life lessons they afford are perhaps more valuable than any technical knowledge I have gained here. Kriti, Mladen, José, Sivaraman, Sourish and many others have enriched my life through countless conversations over the years. Vitor, Einat, Richard, Frank have helped me immeasurably as senior students in helping me learn the ropes. I have also greatly enjoyed discussions with Tae, Bhavana, Mahesh, Dana and many others from William's group.

I would also like to express my immense gratitude for Byron Dom and Dmitry Pavlov, whose mentorship before I came to Carnegie Mellon is responsible for guiding me on my path towards grad school.

Finally, I am indebted to Carnegie Mellon for creating a wonderful environment for intellectual exploration and discovery. The open, friendly atmosphere here has made the campus feel like home and the openness, approachability and generosity of the very accomplished faculty is very humbling. I am also deeply thankful to Stacey Young, Sharon Cavlovich and the rest of the staff at CMU for their invaluable help and assistance over the years.

Abstract

Latent variable mixture models provide an important tool for the analysis of text and relational data. They encompass techniques like topic models for language modeling, and mixed-membership block models, which model relational data that are represented as graphs. A central characteristic of mixed-membership models, is their ability to uncover latent structure from large data in a fully unsupervised manner. Current approaches, however, require us to use these models as black boxes with few avenues to specify the characteristics of the latent structure being uncovered.

In this thesis, we present methods to enable finer control on the characteristics of the inferred latent structure. First, we propose a regularization approach, which is also later placed in a Bayesian framework, that permits the modeler to place preferences on values of aggregate functions over latent variable assignments. This approach is used for instance, to design *slightly mixed membership* models in which entities have only limited freedom to participate in multiple latent roles. Second, we propose methods to introduce limited supervision into the models in form of labeled documents and labeled features. We also introduce a model that performs data fusion between textual and network data to obtain a more robust picture of the underlying latent structure. The advantages of the finer control is demonstrated using an array of text mining tasks such as modeling product reviews, entity clustering and analyses of protein literature and interactions.

Contents

1 Introduction **1**

1.1 Background 1

1.2 Thesis Goal 2

2 Background **5**

2.1 Link-LDA 5

2.1.1 Collapsed Gibbs Sampling for Approximate Inference 8

2.2 Supervised LDA 9

3 Entropic Regularization in Topic Models **11**

3.1 Motivation 11

3.2 Regularizing the Latent Role Distribution of Words 14

3.3 Document Topic Proportion Regularization 18

3.4 Task and Datasets 20

3.5 Experimental Results 22

3.5.1 Effect of Word Entropy Regularization 22

3.5.2 Effect of Document Topic Proportion Regularization 27

3.6 Related Work 32

3.7 Conclusion 35

4 Bayesian Formulation **36**

4.1 Motivation 36

4.2 Entropically Constrained LDA 37

4.2.1 Inference in LDA with ECD Priors . . . 39
4.3 Empirical Results . . . 44
4.4 Related Work . . . 48
4.5 Conclusion . . . 49

5 Limited Supervision in Topic Models 50
5.1 Introduction . . . 50
5.2 Entity Clustering . . . 51
5.3 Exploiting Topic Indicative Features using Regularization based Biased Models . . . 55
5.4 Injecting Labeled Features and Documents . . . 55
5.5 Experimental Results . . . 57
5.6 Related Work . . . 65
5.7 Conclusion . . . 67

6 Entropic Regularization in Network Models 69
6.1 Introduction . . . 69
6.2 Sparse Network Model . . . 70
6.3 Role Entropy Regularization . . . 73
6.4 Cluster Volume Regularization . . . 76
6.5 Experimental Results . . . 78
6.5.1 Datasets . . . 78
6.5.2 Results . . . 80
6.6 Related Work . . . 85
6.7 Conclusion . . . 85

7 Joint Modeling of Network and Documents 86
7.1 Introduction . . . 86
7.2 Block-LDA . . . 88
7.3 Datasets . . . 92
7.4 Experimental Results . . . 95
7.4.1 Results from the Yeast dataset . . . 95

7.4.2 Results from the Enron email corpus dataset 106
7.5 Related work 110
7.6 Conclusion 112

8 Conclusion **113**
8.1 Conclusion 113
8.2 Future Work 114

Bibliography **116**

Chapter 1

Introduction

1.1 Background

Latent variable models have emerged to be a successful and widely used tool to uncover hidden structure in textual corpora. Models like Latent Dirichlet Allocation (LDA) [Blei et al., 2003] treat text documents in a corpus as arising from mixtures of latent topics. In such models, words in a document are potentially generated from different topics using topic specific word distributions. Enabled by the extensible nature of the generative process underlying LDA, many extensions to LDA have been proposed [Erosheva et al., 2004, Griffiths and Steyvers, 2004] which additionally model other metadata, such as authors and tagged entities. Topic models are used for a variety of reasons — as a data exploration tool, for visualization and as a dimensionality reduction technique for a variety of tasks [Andrzejewski and Buttler, 2011, Arora and Ravindran, 2008, Griffiths et al., 2005]. In all these situations, the ability of mixed-membership models to uncover latent structure in data is useful.

Graph analysis has traditionally been dominated by methods based on matrix decompositions. For instance, node clustering, one of the most important applications of graph analysis, is most commonly performed using spectral methods [Luxburg, 2007, Shi and Malik, 2000] which are relaxations of graph cutting methods. More recently, the latent variable mixture model approach that has been widely used for textual corpora has been applied to relational data, i.e., networks. This trend towards stochastic models is similar to the evolution in text models from matrix analysis

methods such as Latent Semantic Indexing to topic models. While empirical studies [Balasubramanyan et al., 2010, Leskovec et al., 2010b] show that flow and graph analysis methods often have computational advantages over latent variable methods, the latter have signficant advantages in terms of modeling flexibility. Latent variable models permit modeling of additional metadata in a more simpler and straightforward manner and provide a greater degree of freedom in modifying models. Stochastic block models [Holland et al., 1983, Snijders and Nowicki, 1997], which were the precursor to mixed-membership block models, aimed to uncover hidden structure in the network by decomposing adjacency matrices of the underlying graph representation of networks into blocks. They posit that nodes in a graph play a single latent role and the probability of an edge depends only on the latent roles of the nodes. While this approach is simple and elegant, nodes in complex graphs often exhibit multiple latent roles. For instance, in a social network, a person might assume a personal role while creating a link with a relative or a family member and don a more professional role while doing the same with a colleague. Airoldi et al. [2008] introduced the mixed membership stochastic block model (MMSB) that models this phenomenon. Parkkinen et al. [2009] later proposed another model which models sparse graphs more efficiently.

Topic models and stochastic block models, while aimed at modeling very different kinds of data, i.e., text corpora and networks, share many attributes. Both classes of models use the idea of mixed-membership where entities — words or documents in the case of text models and nodes in network models, can take on different latent roles everytime they are observed. They also have a notion of topics or clusters which are typically represented as multinomial distributions over nodes or words, thus enabling soft clustering of words.

1.2 Thesis Goal

In this thesis, we focus on the use of mixed-membership models to uncover latent structure in data and address certain shortcomings in this family of models. Current methods in topic modeling and stochastic blockmodeling, for the most part, require that we use them as a "black box" in terms of the latent structure that is uncovered. In general, they lack the facility to enable the modeler to specify the nature of the latent structure that one wishes to extract. While an appealing aspect

of latent variable models is their unsupervised nature, it is often the case that we have a small amount labeled data, which could be advantageous if integrated in the latent structure development. Furthermore, in an era when we are faced with voluminous data from disparate sources, data fusion is critical to ensure that we utilize all the information about the data we are extracting structure from.

The goal of this thesis is to develop modeling techniques that allow modelers to have greater control over the characteristics of the latent structure that is uncovered from data. We present a regularization framework for latent variable models that places preferences on the values of functions over aggregate latent variable assignments. The framework provides considerable flexibility in specifying the characteristics of the latent structure that is inferred, with only limited computational costs. Specifically, we use it to obtain *slightly* mixed membership mixture models where entities are permitted only limited freedom in spanning latent roles. We also present a method to modify the Gibbs sampling approximate inference procedure to incorporated labeled data. These techniques are applied to network models in addition to topic models usually used for text modeling. Finally, we present and evaluate a joint model over networks and text to mine text and related networks.

The rest of the thesis is organized as follows. In chapter 3, we present the regularization framework and demonstrate its advantages in modeling star-annotated product and movie reviews. Using regularized models consistently improve error rates in the task of predicting the star ratings based on the text in reviews. Chapter 4 presents a Bayesian alternate to using the regularizer by introducing a newly proposed prior distribution. We also show empirically that the collapsed Gibbs sampler based inference procedure introduced for the regularized models, also serve as an approximate way to perform inference in a model using the newly proposed prior. In chapter 5, we introduce a method to introduce labels for documents and features into mixed-membership models. Our experiments on clustering entities extracted from web tables show that adding even small amounts of labeled data provide a noticeable improvement in clustering performance. Next, we apply the regularization framework proposed in chapter 3 to network block models in chapter 6. We use the regularized network models to perform node clustering in graphs and show that regularization provides improvements in node cluster recovery. Finally, we present a model to perform data fusion by jointly modeling networks and documents in chapter 7. Our final conclusions

are presented in chapter 8.

Chapter 2

Background

In this chapter, we present background information on the mixed-membership latent variable models that are used as the foundation for the rest of the thesis. We specifically describe the Link-LDA [Erosheva et al., 2004, Nallapati et al., 2008] and supervised LDA [Blei and McAuliffe, 2008] models. These models are based on Latent Dirichlet Allocation (LDA) [Blei et al., 2003] which relaxes the more typical clustering assumption that every document is generated from a single underlying topic in what we refer to as the Mixture of Multinomials (MoM) model. Instead, in LDA-based models, each document has an unique distribution of underlying topics and is typically drawn from a Dirichlet prior.

2.1 Link-LDA

The Link-LDA model (Figure 2.1) [Erosheva et al., 2004, Nallapati et al., 2008] is an extension of Latent Dirichlet Allocation [Blei et al., 2003] where documents are augmented with a bag of references in addition to a bag of words. The bag of words [Salton and McGill, 1986] representation is a widely used model for documents in text modeling. In this representation, every word in a document is considered to be exchangeable with other words in the document, therefore causing word order to be ignored. In the Link-LDA model, the notion of bag of words is extended and documents are treated as a set of bag of words. Documents are represented as containing T types of entities, i.e. they are represented as T "bags-of-words". When $T = 1$, the model reduces to

γ_t - Symmetric Dirichlet prior hyperparameter for topic multinomials of type t
$\beta_{t,k}$ - multinomial over entities of of type t in the vocabulary V_t for topic k
α - Symmetric Dirichlet prior hyperparameter for document specific topic distributions
θ_d - multinomial distribution over K topics for document d
$z_{t,i}$ - topic chosen for the i-th entity of type t in a document, $z_{t,i} \in 1, \ldots, K$
$w_{t,i}$ - the i-th entity of type t occurring in a document, $w_{t,i} \in 1, \ldots, V_t$

Table 2.1: Parameters in LDA-like models

LDA.

In Link-LDA, a corpus of documents D is modeled using parameters listed in Table 2.1. Each entity type has a topic wise multinomial distribution over the set of entities that can occur as an instance of the entity type. For instance while modeling scientific literature in machine learning, the types of entities could be words, metadata that the papers have been tagged with like authors, algorithms, datasets, and metrics. This extended topic model can therefore effectively model documents containing different sets of entities. It is important to note that words in a document are treated in the same manner as other kinds of entities, i.e. they are simply a particular entity type. The generative process underlying the model is as follows:

1. Generate topics: sample $\beta_{t,k} \sim \text{Dir}(\gamma_t)$ for $t \in 1, \ldots, T$, $k \in 1, \ldots, K$
2. Generate documents: For each document $d \in D$
   (a) Sample $\theta_d \sim \text{Dir}(\alpha)$
   (b) For each type of entity $t \in \{1, \ldots, T\}$
      i. For each instance of an entity $w_{t,i}, i \in \{1, \ldots, N_{d,t}\}$
         A. Sample a topic $z_{t,i} \sim \text{Multinomial}(\theta_d)$
         B. Sample $w_{t,i}$ from $\beta_{t,z_{t,i}}$

The model is parameterized by $\boldsymbol{\beta_{1\ldots T,1\ldots K}}$, i.e., a multinomial for every entity type and every topic.

To represent documents in the latent topic space, we need to run inference and compute the posterior distributions over topics for each document. Exact inference on the LDA and Link-LDA

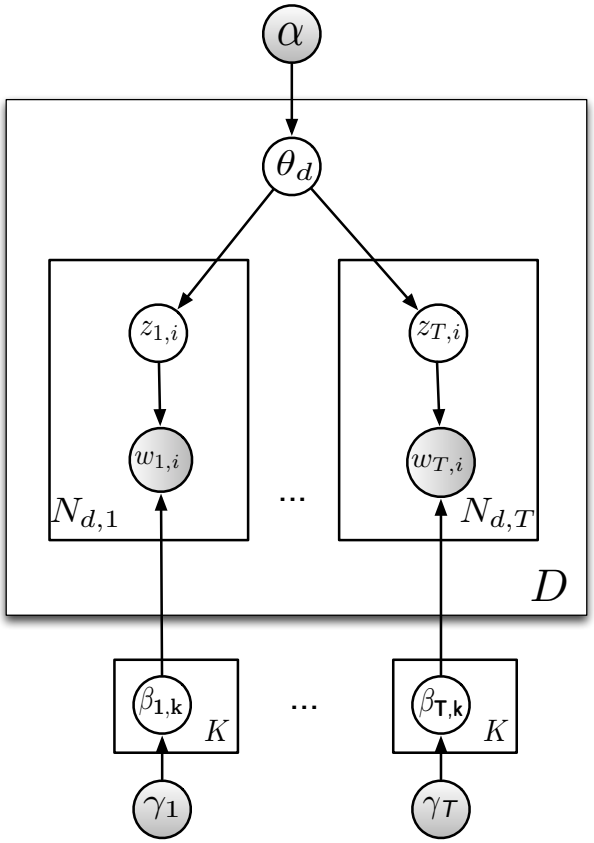


Figure 2.1: The Link-LDA model

models is however not tractable. Approximate inference techniques, of which there are many in the literature, are therefore used. Variational inference techniques [Blei et al., 2003, Ghahramani and Beal, 2000] are one popular family of algorithms used for this purpose. In this approach, the likelihood function is bounded by a simpler likelihood function that eliminates dependencies that make inference intractable. The inference procedure proceeds by fitting values to the simplified model whilst ensuring that the KL divergence to the original model is kept low. A second popular family of methods that is frequently used, is based on Markov Chain Monte Carlo (MCMC) methods. When conditional probabilities of the variables for which samples are needed can be expressed easily, Gibbs sampling, which is a specific type of MCMC method, can be used. For more complex models, more computationally expensive forms of MCMC methods such as Metropolis Hastings are

also used. In this thesis, we primarily use collapsed Gibbs sampling [Griffiths and Steyvers, 2004, Porteous et al., 2008] for approximate inference. Asuncion et al. [2009] provides an overview of different inference techniques and shows that the commonly used methods are ultimately similar when the update equations used for the techniques are considered.

2.1.1 Collapsed Gibbs Sampling for Approximate Inference

MCMC methods can emulate high-dimensional probability distributions by the stationary behaviour of a Markov chain. This means that one sample is generated for each transition in the chain after a stationary state of the chain has been reached, which happens after a so-called "burn-in period" that eliminates the influence of initialisation parameters. Gibbs sampling is a special case of MCMC where the random variables of the distribution are sampled alternately one at a time, conditioned on the values of all other variables. The algorithm proceeds by choosing a random variable i and sampling x_i conditioned on all other variable.

Here, we discuss the collapsed Gibbs sampling expression for the Link-LDA model [Griffiths and Steyvers, 2004]. The joint distribution of the Link-LDA model is defined as:

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w} | \alpha, \boldsymbol{\gamma}) = \prod_{k=1}^{K} \prod_{t} \text{Dir}(\beta_{t,k} | \gamma_t) \left(\prod_{d=1}^{D} \text{Dir}(\theta_d | \alpha) \left(\prod_{t} \left(\prod_{i=1}^{N_{d,t}} \theta_d^{(z_{t,i})} \beta_{t,z_{t,i}}^{(w_{t,i})} \right) \right) \right) \tag{2.1}$$

Following the derivation in Heinrich [2009], the sampling equation i.e., the conditional probability of sampling a value for $z_{t,i}$ conditioned on assignments to all other z indicators, after collapsing θ and $\boldsymbol{\beta}$, is given by:

$$p(z_{t,i} = k | w_{t,i} = w, \mathbf{z}^{\neg \mathbf{i}}, \mathbf{w}^{\neg \mathbf{i}}, \alpha, \boldsymbol{\gamma}) \propto \frac{\sum_{t'} n_{dkt'}^{\neg i} + \alpha}{\sum_{t'} \sum_{k'} n_{dk't'}^{\neg i} + K\alpha} \frac{n_{kwt}^{\neg i} + \gamma_t}{\sum_{w'} n_{kw't'}^{\neg i} + |V_t|\gamma_t} \tag{2.2}$$

The n's refer to number of topic assignments in the data.

- n_{kwt} - the number of times an entity w of type t is observed under topic k
- n_{dkt} - the number of entities of type t with topic k in document d

The superscript $\neg i$ for the counts indicates that the current word for which the topic indicator is being sampled is excluded from the counts. Similary $\mathbf{z}^{\neg \mathbf{i}}$ in Equation 2.2 indicates all the z variables in the document except the one for the word currently under consideration.

The inference procedure proceeds by cyclically sampling values for each $z_{t,i}$ variable using the distribution defined in equation 2.2. The procedure is repeated for several iterations where an iteration is defined as the process of obtaining samples for each z variable. The number of iterations used is either manually set to a sufficiently high number or is determined to be the number of iterations required for the perplexity of the dataset to be relatively stable.

Further details about collapsed Gibbs sampling in LDA-like models can be found in Newman et al. [2006a].

2.2 Supervised LDA

Supervised LDA (SLDA) introduced in Blei and McAuliffe [2008] extends LDA to model documents with response variables. The model is designed to uncover topics that serve a dual purpose — to explain the contents of documents and to predict the response variables. In SLDA, the response variables are generated using a generalized linear model. For the rest of the thesis, we use a normal linear model to produce responses using the topic proportions for the document as the covariates. The plate diagram for the model is given in Figure 2.2.

The generative process of the model is as follows:

1. Sample $\beta_k \sim \text{Dir}(\gamma)$ for $k \in 1, \ldots, K$
2. For each document d
   (a) Sample $\theta_d \sim \text{Dir}(\alpha)$
   (b) For each instance of a word $w_i, i \in \{1, \ldots, N_d\}$
      i. Sample a topic $z_i \sim \text{Multinomial}(\theta_d)$
      ii. Sample w_i from β_{z_i}
   (c) Draw response variable y_d from $\mathcal{N}(\overline{\mathbf{z}_d}\eta, \sigma^2)$

Note that $\overline{\mathbf{z}_d}$ represents the smoothed topic proportion distribution of the document and is

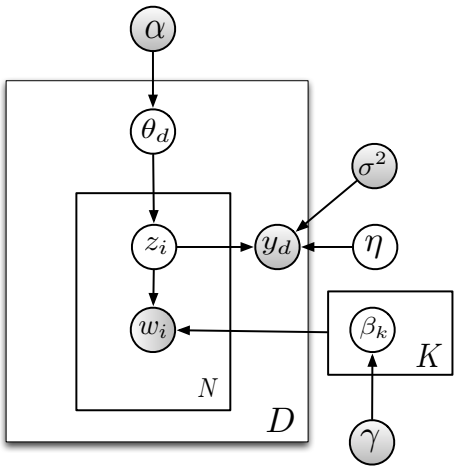


Figure 2.2: Supervised LDA (sLDA)

given by $\overline{z_d}^{(k)} = \frac{\sum_{i=1}^{N_d} \mathbf{I}(z_i=k)+\alpha}{N_d+K\alpha}$.

As described above, the SLDA model includes a regression model and its associated parameters i.e., η and σ^2, that need to be fit in addition to the usual LDA parameters. Therefore the collapsed Gibbs sampling procedure is extended to alternate between sampling values for the z, the topic indicators and training the regression model. After every iteration of the collapsed Gibbs sampler, the sampled values for z are used to provide training data to fit the regression model and obtain estimates for η and σ^2.

The sampling equation for sampling topics, after collapsing θ and $\boldsymbol{\beta}$ is based on the collapsed Gibbs sampler for Link-LDA and is given by:

$$\begin{aligned} & p(z_i = k | w_i = w, \mathbf{z}^{\neg \mathbf{i}}, \mathbf{w}^{\neg \mathbf{i}}, \alpha, \boldsymbol{\gamma}) \\ & \propto \frac{n_{dk'}^{\neg i} + \alpha}{\sum_{k^i} n_{dk'}^{\neg i} + K\alpha} \frac{n_{kw}^{\neg i} + \gamma}{\sum_{w'} n_{kw'}^{\neg i} + |V|\gamma} \mathcal{N}(\eta \overline{\mathbf{z}_d}, \sigma^2) \end{aligned} \tag{2.3}$$

Similar to the counts defined for Link-LDA previously, the n's refer to number of topic assignments in the data.

- n_{kw} - the number of times a word w is observed under topic k
- n_{dk} - the number of words with topic assignment k in document d

Details on the derivation of the sampling equation can be found in Chang [2011].

Chapter 3

Entropic Regularization in Topic Models

3.1 Motivation

In topic models based on LDA [Blei et al., 2003], when a word is observed multiple times in a corpus, the latent topics that generate the different word instances are not constrained to be the same for every instance. This useful property of LDA is instrumental in modeling natural language phenomena like *polysemy* and *homonymy*, i.e., the capacity of a word to have several different senses. For instance, we typically expect the model to assign different latent roles to occurrences of a polysemous word like *bank*, allowing it to take part in different topics (e.g., *finance* and *fishing*). The freedom of a word to participate in multiple topics can be contrasted with hard clustering based approaches like Brown clustering [Momtazi and Klakow, 2009] . While the freedom granted by LDA to permit a word to take on multiple latent roles is useful, this flexibility can be overly permissive. For instance, topic models often contain hundreds of topics, whereas the average number of senses for a word on average is far lower. Marquez et al. [2006] report that the range of the number of word senses for an ambiguous set of words in the DSO [Ng and Lee, 1997] corpus was 3 to 25 with a mean of 10.1. In general we anticipate that most words will have few meanings, and hence the set of latent topics associated with most words will be small. While it is often empirically observed that the posterior latent role distributions of words are sparse, especially when a sparse prior is employed, our experiments in Section 3.5 show that the inference procedure still returns posteriors that span a wider range of topics than is optimal.

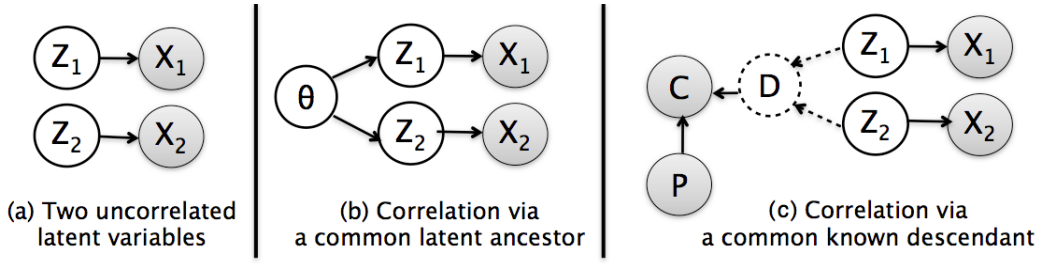


Figure 3.1: Different ways to impose a correlation between latent variables.

The description above regarding the overly permissive nature of topic models therefore necessitates a method to model *slightly* mixed-membership in words. This requirement is often made more apparent when LDA is used to model non-textual data. When one is modeling such data, the common approach is to define analogues between words and images in LDA with their counterparts in the domain from which the data is drawn. For instance, in the modeling of image data [Cai et al., 2008], images are treated analogously to documents and are represented as a bag of "visual words". "Words" in such situations could potentially exhibit multiple senses to a lesser degree than words in natural language, making the need to model slightly mixed-membership more acute.

Rubin et al. [2012] also show that the number of labels that are assigned to documents in a multi-label framework is much lower than the number of topics typically used in topic modeling. It can therefore be advantageous to build models that are coaxed to activate fewer topics to model a document.

Here, we present a regularization framework that is designed to control the freedom that LDA-like models accord to words. The primary requirement of such a framework is to introduce correlations between words' probabilities in different topics that are iid in an unregularized model. Generally, when designing probabilistic models, a modeler has two options to ensure correlations between latent variables: including a common latent ancestor of the latent nodes, or including a common observed descendant. Figure 3.1 shows a very simple example of this modeling choice, in the case of a Bayes network with four binary variables: for the sake of concreteness, suppose that Z_1 and Z_2 represent some genetic property, and X_1 and X_2 encode an observable phenotype (say baldness). In part (a) of the figure the latent variables Z_1 and Z_2 are independent. In part (b) the variables are dependent, by virtue of the new node θ.

The model of (b) is quite useful if θ has a plausible causal interpretation—here, e.g., X_1 and X_2 might be observations of two brothers, θ genetic properties of their father, and $P(Z_i|\theta)$ a model of inheritance. In some settings, however, the causal interpretation is either poorly understood or difficult to model, but information is available about correlation. For instance, we may know that one in four men is bald, but that one pair of brothers in eight is also bald (rather than the expected one in 16). We can model this information conveniently by explicitly introducing variables that measure such correlations, and then using priors to express our preferences over them. As shown in Figure 3.1 part (c), we can specify D to be true iff both X_1 and X_2 are true and let C be a noisy copy of D whose expectation is encoded by the known variable P (which acts as a prior) and $P(C|P)$. (In this case, D is a deterministic function of its parents, so it is conceptually awkward to introduce a prior directly on D). Approach (c) also makes Z_1 and Z_2 dependent, and in this simple case, it is easy to show that the same set of probability distributions can be modeled by the networks in (b) and (c); however, the approach has different implications computationally.

Modeling choice (c) can be generalized as follows: given an existing generative model, we can bias it by introducing new variables that, like D, measure some aggregate property of the latent variables, and then imposing a prior preference over this aggregate value. One computationally convenient way to introduce such a prior is to allow the aggregate node D to be latent, but introduce a new node C that is a "noisy copy" of C, with a specified noise model.

We show that we can use this technique to encourage lower entropy in the latent topic assignments to a particular word. Perhaps surprisingly, these bias variables can be introduced with minimal computational change: for instance, they can be added to topic models without impacting conjugacy. Importantly, they can be conveniently added to modify aggregates that are never explicitly sampled, and hence are difficult to predictably modify by simply changing the priors of a model. In the proposed framework, the model is extended to include a noisy copy of an aggregate function over latent variables, such as the entropy of the topic distributions. By pretending to see (*pseudo-observing*) a desired value for the copy the model is coaxed to push the variables that participate in the aggregate functions to values that make the pseudo-observed values likely. The aggregate values depend on latent variable values and pseudo-observed variables in ways that are not possible otherwise. For instance in topic models based on LDA, as described earlier, differ-

ent topics have the freedom to generate the same word. Since the topics are drawn iid from a (typically conjugate) prior, there is no direct way to create a dependency between the probability of generating a given word and the probability of another topic generating the same word. The regularization framework proposed here permits us to overcome this restriction by crafting suitable aggregate functions without requiring complicated priors that prevent us from collapsing θ and $\boldsymbol{\beta}$. The approach also has the advantage of keeping posterior inference in the model simple; we need only extend the commonly used collapsed Gibbs sampler used for approximate inference in latent variable models, by adding a few additional terms. The addition of these terms can result in no increase in computational order of complexity of inference if functions are chosen such that their values can be cached and updated with $O(1)$ cost when latent roles assignments for a word changes.

It should be noted that the proposed regularization framework can be used to place preferences on any function that operates on the latent variable assignments of words. By designing suitable functions as described later in the chapter, preferences can also be placed on the properties of a documents' topic distribution. Here, we focus on using it to control the entropy of the latent role distribution of words and documents and leave other uses of the framework for future work.

3.2 Regularizing the Latent Role Distribution of Words

As explained earlier (Section 3.1), it is often desirable to explicitly bias LDA-like models to prefer lower entropies in words' latent role distributions. Here, we describe how we use the modeling approach described above to incorporate such constraints into the SLDA model. The scheme is illustrated in Figure 3.2. As described in the Chapter 2.2, SLDA extends LDA by incorporating response variables in documents that are dependent on their topic proportions. The inference procedure for SLDA jointly learns parameters for the topic model and the regression model. The regularization framework can be added to any LDA-derived model; we use the document-level measurement modeling property of SLDA here to predict star ratings for movie reviews. We evaluate the method by computing the average squared error in predicting the rating scores. The entropy based regularization technique based on pseudo-observed variables directly controls the

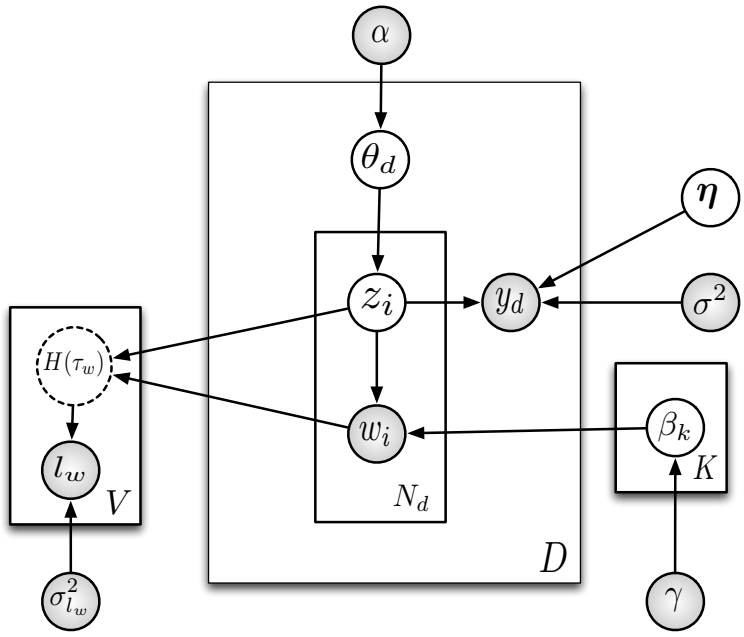


Figure 3.2: Supervised LDA with Word Latent Role Distribution Regularization.

freedom of words to take on different latent topics, by penalizing high entropies in their topic distributions.

First, we formally define the latent role distribution of a word, τ_w. A point estimate for the topic distribution of a word $w \in V$ (i.e. the vocabulary) in SLDA can be defined as

$$\tau_w^{(k)} = \frac{n_{kw}}{\sum_{k'} n_{k'w}},\ k \in 1, \ldots, K \tag{3.1}$$

The n's refer to number of topic assignments in the data, as follows:

n_{kw} - the number of times a word w is observed under topic k

n_{dk} - the number of entities with topic k in document d

Note that K refers to the number of topics in the model. The Shannon entropy of this distribution can be computed using the expression $H(\tau_w) = -\sum_k \tau_w^{(k)} \log_2 \tau_w^{(k)}$.

In essence, $H(\tau_w)$ indicates how "diffuse" a word is across different topics. A high entropy value indicates that the word has nearly equal probability of being generated by different topics, whereas a low value indicates that the word has a high chance of being generated by a few topics and low probability of being generated by the rest of the topics.

As seen in the plate diagram, we now introduce word topic distribution entropy regularization by adding pseudo-observed variables, l_w (Figure 3.2), one for each word in V, which can be interpreted as noisy copies of $H(\tau_w)$. In the figure the $H(\tau_w)$ nodes are shown in a dotted circle since these are variables deterministically determined from the latent role assignments.

To enforce the preference for entropy values as specified by l_w, the generative story specifies that these copies are drawn from a truncated one-sided normal distribution parameterized by a mean of $H(\tau_w)$ and a variance hyperparameter $\sigma^2_{l_w}$. The probability of observing the noisy copy is therefore given by

$$p(l_w|h = H(\tau_w), \sigma^2_{l_w}) = \frac{1}{C} \exp\left(\frac{-(l_w - h)^2}{2\sigma^2_{l_w}}\right), 0 \leq l_w \leq \log_2 K \tag{3.2}$$
$$\text{where } C = \int_{h'=0}^{\log_2 K} \exp\left(\frac{-(h' - l_w)^2}{2\sigma^2_{l_{t,w}}}\right) dh.$$

The addition of the regularization terms with l_w set to 0 (or any other low value), penalizes large entropies in the topic distributions of words, with $\sigma^2_{l_w}$ dictating the strictness of the penalty. The penalization of large entropies therefore drives the inference procedure to return models that exhibit lower entropies in their word distributions. In effect, the inference procedure balances the need for the topic distributions to fit the observed words and the need to restrict words' latent role distribution entropies. As stated earlier, to generalize the framework, $H(\tau_w)$ can be substituted with any arbitrary function that operates over the latent and observed variables and l_w can be correspondingly set to a preferred value for the function. The addition of these terms have an impact on computational efficiency during collapsed Gibbs sampling. By choosing functions that can be implemented efficiently, we can introduce the regularization with no difference in the computational order of complexity of Gibbs sampling.

The joint distribution of the SLDA model with regularization is defined as:

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w}, \mathbf{y}, \mathbf{l_w}|\alpha, \gamma, \boldsymbol{\eta}, \sigma^2, \sigma^2_{l_w}) = \tag{3.3}$$
$$\prod_{k=1}^{K} \text{Dir}(\beta_k|\gamma) \left(\prod_{d=1}^{D} \text{Dir}(\theta_d|\alpha) \left(\prod_{i=1}^{N_d} \theta_d^{(z_i)} \beta_{z_i}^{(w_i)}\right)\right) \left(\prod_{w \in V} \exp \frac{-\left(l_w - H(\tau_w)\right)^2}{2\sigma^2_{l_w}}\right) \exp\left(\frac{-(y_d - \boldsymbol{\eta}.\overline{\mathbf{z}_d})^2}{2\sigma^2}\right)$$

Note that $\overline{\mathbf{z}_d}$ represents the smoothed topic proportion distribution of the document and is given by $\overline{z_d}^{(k)} = \frac{\sum_{i=1}^{N_d} \mathbf{I}(z_i = k) + \alpha}{N_d + K\alpha}$.

Next, we derive the collapsed Gibbs sampling equation to sample a topic indicator for a word w_i for the regularized model

$$p(z_i = k|\mathbf{z}^{\neg\mathbf{i}}, \mathbf{w}, y_d, \mathbf{l_w}, \alpha, \gamma, \boldsymbol{\eta}, \sigma^2, \sigma^2_{l_w}) \propto p(z_i = k|\mathbf{z}^{\neg\mathbf{i}}, \mathbf{w}, \alpha, \gamma)\ \ p(y_d|\overline{\mathbf{z}_d}, \boldsymbol{\eta}, \sigma^2)\ \ p(l_{w_i}|\tau_{w_i}, \sigma^2_{l_w}) \tag{3.4}$$

Using $p(z_i = k|\mathbf{z}^{\neg\mathbf{i}}, \mathbf{w}, \alpha, \gamma) \propto (n^{\neg i}_{dk} + \alpha)\frac{n^{\neg i}_{kw_i} + \gamma}{\sum_{w'} n^{\neg i}_{kw'} + |V|\gamma}$ from Equation 2.3, we get

$$\begin{aligned} &p(z_i = k|\mathbf{z}^{\neg\mathbf{i}}, \mathbf{w}, y_d, \mathbf{l_w}, \alpha, \gamma, \boldsymbol{\eta}, \sigma^2, \sigma^2_{l_w}) \propto \\ &(n^{\neg i}_{dk} + \alpha)\frac{n^{\neg i}_{kw_i} + \gamma}{\sum_{w'} n^{\neg i}_{kw'} + |V|\gamma}\exp\left(\frac{-(y_d - \boldsymbol{\eta}.\overline{\mathbf{z}_d})^2}{2\sigma^2}\right)\exp\left(\frac{-(H(\tau_{w_i}) - l_{w_i})^2}{2\sigma^2_{l_w}}\right) \end{aligned} \tag{3.5}$$

The sampling equation for regularized SLDA (Eqn 3.5) is therefore derived by incorporating the additional terms added to the SLDA model for regularization into the sampler. It should be noted that the terms $H(\tau_{w,i})$ and $\overline{\mathbf{z}_d}$ are computed using the assignment $z_i = k$, i.e., the assignment to word i is not excluded unlike in the n counts. The computational complexity of collapsed Gibbs sampling is O(IN_dK), where I is the number of iterations. Note that $H(\tau_w)$ can be computed in O(1) time by caching $n_{kw} \log_2 n_{kw}$ values. Similarly $\boldsymbol{\eta}.\overline{\mathbf{z}_d}$ can also be computed in O(1) by caching the product and adjusting it whenever a word's latent role assignment is changed.

As we will see in the experiments below, during the collapsed Gibbs sampling process, the inference procedure tends to push $H(\tau_w)$ close to the *pseudo-observed* l_w. Setting l_w to 0 therefore coaxes the inference procedure to return low entropy topic distributions for the words in the vocabulary. The variance parameter $\sigma^2_{l_w}$ can be used to adjust the tightness of the Gaussian to permit more or less entropy in the label distributions.

We can also see that the framework does not require $\sigma^2_{l_w}$ to be the same for all words or even that the regularization be applied to all the words in the vocabulary. If a modeler so requires, regularization can be restricted to a subset of the vocabulary and can be applied with different variances to different words.

An alternate method to achieve such entropic sparsity is to modify the Dirichlet priors with a different prior distribution that prefers topics with similar low word latent role distribution entropy properties. Doing so however requires complicated priors (which can of course no longer be Dirichlet) that are capable of producing topic distributions that are not iid, leading to complications in

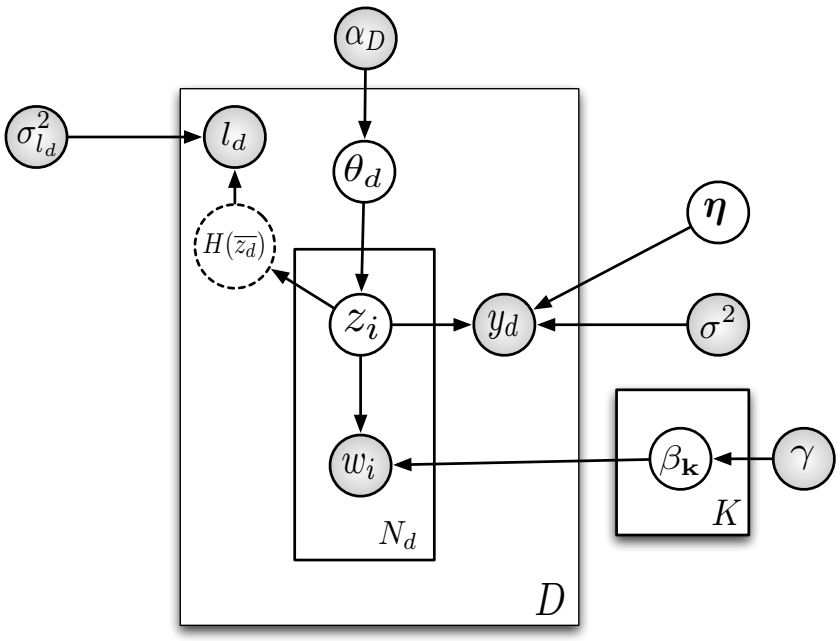


Figure 3.3: Supervised LDA with Document Topic Regularization.

the inference procedure. Chapter 4 discusses this approach in detail. The regularization technique described here results in a similar effect on word entropy distributions. It requires minimal additions to the existing collapsed Gibbs sampling inference procedure but is arguably less elegant than using a different prior due to the loss of intuitiveness of the generative process.

3.3 Document Topic Proportion Regularization

In models based on LDA, the topic proportions of documents are typically drawn from a Dirichlet distribution. It is commonly seen in practice that the posteriors of these distributions obtained after inference are often sparse. In this section, we use the entropic regularization technique introduced earlier in the chapter to explicitly increase and control the sparsity in the topic proportion distribution.

For every document d in the corpus, we define $H(\overline{\mathbf{z}_d})$ as the Shannon entropy of the observed topic proportion distribution (See Figure 3.3). Since $\overline{\mathbf{z}_d}$ is obtained by repeatedly sampling θ_d, it

can be considered as an approximate point estimate of θ_d. We use the same framework as the previous section to explicitly coerce the inference procedure to return low entropy topic proportion distributions. To incorporate the regularization, we add a noisy copy of the entropy score, l_d, which is sampled from a distribution that is parameterized by $H(\overline{\mathbf{z}_d})$ and a hyperparameter $\sigma^2_{l_d}$. The density function for l_d is the same as specified in Equation 3.2. $H(\overline{\mathbf{z}_d})$ is again shown in a dotted circle since it is deterministically computed from the latent role assignments to words in document d. Setting l_d to a low value forces the inference procedure to favor a θ_d that has a low entropy value as well.

The joint likelihood of the model with the document topic proportion regularization added is given by

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w}, \mathbf{y}, \mathbf{l_d} | \alpha, \gamma, \boldsymbol{\eta}, \sigma^2, \sigma^2_{l_d}) \propto \tag{3.6}$$
$$\prod_{k=1}^{K} \text{Dir}\left(\beta_k | \gamma\right) \left(\prod_{d=1}^{D} \text{Dir}(\theta_d | \alpha) \left(\prod_{i=1}^{N_d} \theta_d^{(z_i)} \beta_{z_i}^{(w_i)} \right) \exp \frac{-\left(l_d - H(\overline{\mathbf{z}_d})\right)^2}{2\sigma^2_{l_d}} \right)$$

The collapsed Gibbs sampling condtional distribution to sample a topic indicator for a word can be derived in a similar manner to the derivation of Equation 3.5 and is given by:

$$p(z_i = k | \mathbf{l_w}, l_d, w_i, \mathbf{z}^{\neg \mathbf{i}}, \mathbf{w}^{\neg \mathbf{i}}, y_d, \alpha, \gamma, \boldsymbol{\eta}, \sigma^2, \sigma^2_{l_d}) \propto \tag{3.7}$$
$$(n_{dk}^{\neg i} + \alpha) \frac{n_{kw_i}^{\neg i} + \gamma}{\sum_{w'} n_{kw'}^{\neg i} + |V|\gamma} \exp\left(\frac{-(y_d - \boldsymbol{\eta}.\overline{\mathbf{z}_d})^2}{2\sigma^2} \right) \exp\left(\frac{-(l_d - H(\overline{\mathbf{z}_d}))^2}{2\sigma^2_{l_d}} \right)$$

If both forms of regularization were to be used, the likelihood of data is given by the expression

$$\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{z}, \mathbf{w}, \mathbf{y}, \mathbf{l_w}, \mathbf{l_d} | \alpha, \gamma, \boldsymbol{\eta}, \sigma^2, \sigma^2_{l_w}, \sigma^2_{l_d}) = \tag{3.8}$$
$$\prod_{k=1}^{K} \text{Dir}(\beta_k | \gamma) \left(\prod_{d=1}^{D} \text{Dir}(\theta_d | \alpha) \left(\prod_{i=1}^{N_d} \theta_d^{(z_i)} \beta_{z_i}^{(w_i)} \right) \exp \frac{-\left(l_d - H(\overline{\mathbf{z}_d})\right)^2}{2\sigma^2_{l_d}} \right) \times \exp\left(\frac{-(y_d - \boldsymbol{\eta}.\overline{\mathbf{z}_d})^2}{2\sigma^2} \right)$$
$$\left(\prod_{w \in V} \exp \frac{-\left(l_w - H(\tau_w)\right)^2}{2\sigma^2_{l_w}} \right) \tag{3.9}$$

The conditional distribution for collapsed sampling equation is simply extended to include both

Dataset	books	DVD	kitchen	electronics	movies
Range	1 - 5	1 - 5	1 - 5	1 - 5	0 - 1
Size	5501	5118	5149	5901	2000
Vocabulary	13743	14548	10857	8377	12639

Table 3.1: Dataset statistics.

regularization terms and is defined as:

$$p(z_i = k|\mathbf{l_w}, \mathbf{l_d}, w_i, \mathbf{z}^{\neg \mathbf{i}}, \mathbf{w}^{\neg \mathbf{i}}, y_d, \alpha, \gamma, \boldsymbol{\eta}, \sigma^2, \sigma^2_{l_d}, \sigma^2_{l_w}) \propto \tag{3.10}$$
$$(n^{\neg i}_{dk} + \alpha)\frac{n^{\neg i}_{kw_i} + \gamma}{\sum_{w'} n^{\neg i}_{kw'} + |V|\gamma}\exp\left(\frac{-(y_d - \boldsymbol{\eta}.\overline{\mathbf{z}_d})^2}{2\sigma^2}\right)\exp\left(\frac{-(l_d - H(\overline{\mathbf{z}_d}))^2}{2\sigma^2_{l_d}}\right)\exp\left(\frac{-(H(\tau_{w_i}) - l_{w_i})^2}{2\sigma^2_{l_w}}\right)$$

As $\sigma^2_{l_d}$ tends to 0, the model reduces to a mixture of multinomials model since the regularization will require the entropies to be close to 0 implying that the distribution over topics has all its mass on one topic. Similarly, as the variance tends to ∞, the model reduces to a fully unconstrained LDA-like model.

Unlike word latent role distribution regularization, where the distribution is not explicitly sampled, θ_d is explicitly sampled from a Dirichlet prior. Therefore, using a small symmetric Dirichlet hyperparameter value can provide sparse distributions which have low entropies. We present a comparison of the regularization approach and the hyperparameter tuning approach in section 3.5.2.

3.4 Task and Datasets

The regularization approach presented can be used in any situation where LDA-derived models are used. Here we use the regularization to improve the performance on the SLDA model in predicting sentiment from product reviews. Automatically discerning the sentiment expressed in reviews of products, hotels and movies has been an active area of research [Jo and Oh, 2011, Joshi et al., 2012, Pang and Lee, 2005, Titov and McDonald, 2008a]. Here we specifically look at the task of predicting the star-rating score typically seen in online reviews (ranging between 1-5 for product reviews on Amazon.com) based on the text in the product reviews. This task was previously tackled in Titov and McDonald [2008b] where the authors proposed a multi-grain topic model to extract

ratings from hotel reviews, and achieve a best absolute error of 0.669. Qu et al. [2010] use a variant of ridge regression to get mean squared errors of $0.884, 0.928$ and 0.627 in predicting the scores of product reviews (a different set of products than used by us) in the book, DVD and music domains respectively. Following the precedent set in previous work in the field, we employ the average mean squared error (MSE) in predicting star-rating scores as the metric of evaluation

Here, we test the performance of an entropy regularized SLDA model by modeling star-rating annotated reviews of products from Amazon.com. The datasets provided by Blitzer et al. [2007][1] consist of 4 sets of reviews corresponding to reviews about products in the books, DVD, electronics and kitchen categories. Each review is annotated with a star-rating that ranges from 1 to 5. We use only the text of reviews as inputs in this study. We also test the model on a movie reviews dataset [Pang and Lee, 2005] that contains 2000 reviews labeled as positive or negative, which are tagged with targets 1 and 0 respectively in our experiments. Statistics about the datasets are shown in Table 3.1.

Experimental Setup

First, we present an overview of the approach. As the first step, we train a SLDA model using the text of product reviews and their associated star ratings. Review documents are represented as a bag of unigram features. After the model is trained, when a previously unseen review is encountered during test time, we use the trained model to perform inference and get a topic distribution for the review. The trained regression model in SLDA is then used to map the topic distribution to a predicted star-rating value. For all the experiments in the rest of the chapter, we set the pseudo observed entropy value to be 0 and set the variance parameter $\sigma^2_{l_w}$ to different values to test the sensitivity of the model to the hyperparameter. The topic distributions $\boldsymbol{\beta}$, document topic proportions $\boldsymbol{\theta}$ and the regression parameters are estimated during inference using the collapsed Gibbs sampler described earlier. The Dirichlet hyperparameter values α and γ are set to 0.2 and are not further optimized. Inference on the SLDA model is performed using a collapsed Gibbs

[1] http://www.cs.jhu.edu/∼mdredze/datasets/sentiment/

sampler where we let the sampler burn-in for 100 iterations[2] and take the average of the 10 samples after burn-in for obtaining estimates for the model parameters. Unless stated otherwise, all results presented are averaged over 10 trials starting with different randomly chosen start states for the sampler. Error bars in the graphs show the variance in performance across different trials. In each trial, results presented on the test set are obtained through 10-fold cross validation.

3.5 Experimental Results

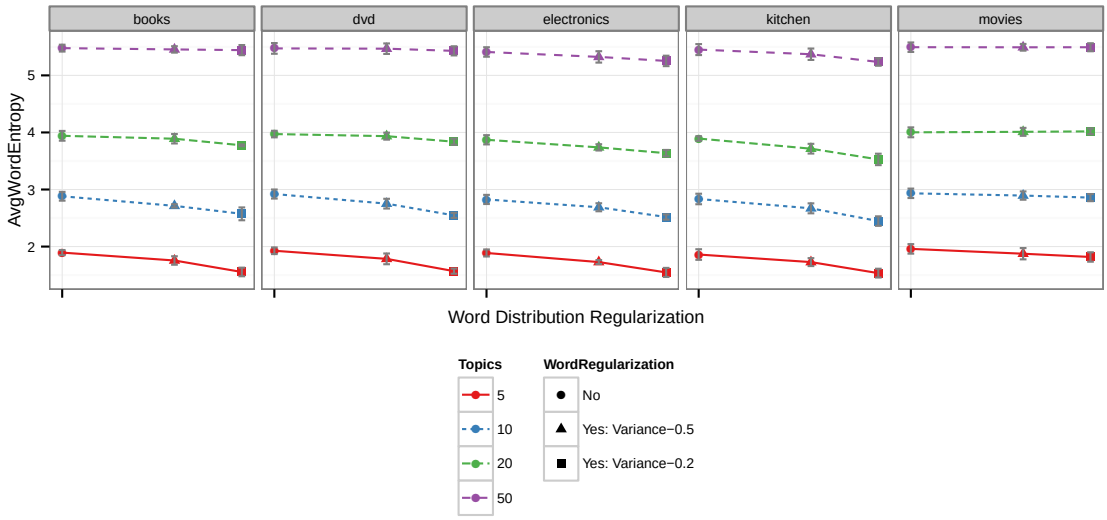
3.5.1 Effect of Word Entropy Regularization

We first investigate the effect of the proposed regularization on the entropy of words' topic distributions by studying the change in the word topic distribution entropy, averaged across all words in the vocabulary, i.e. $\sum_{w \in V} H(\tau_w)/|V|$. Figure 3.4(a) shows the change in the average word topic entropy for the datasets described above as the regularization is applied and increasingly tightened by decreasing the value of the variance parameter $\sigma_{l_w}^2$. The different lines in the plots indicate results of trials with different values of K (the number of topics). The first point in each plot indicates results with no regularization applied. The second and third points in each line show the entropy values with the variance value set to 0.5 and 0.2 respectively.

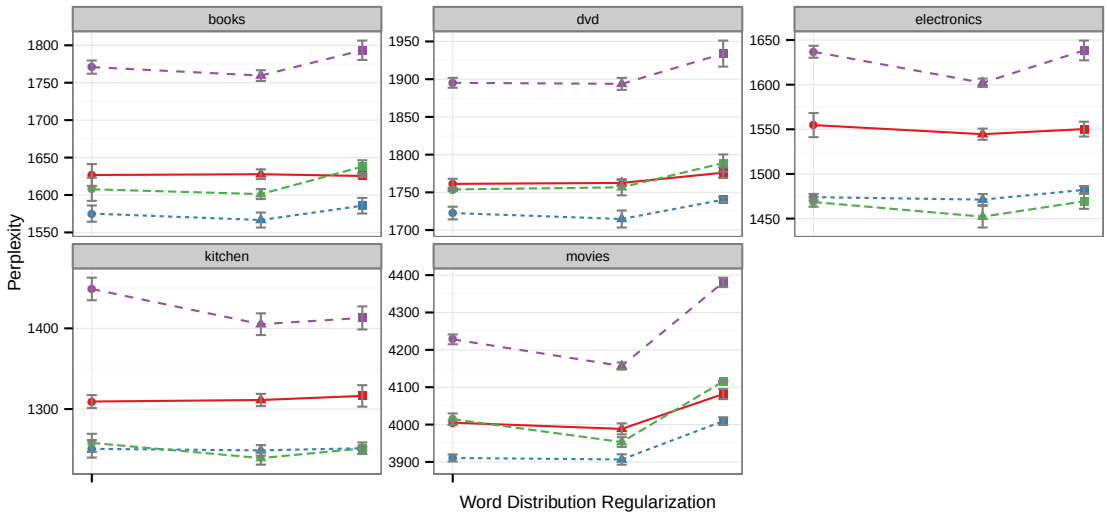
First, it can be seen that the models with higher K show higher absolute entropy values, since words have a greater number of topic indicators that they can appear under in. All the plots show a general downward trend in the average word entropy as regularization is applied and the regularization hyperparameter variance is decreased. This is explained due to the heavier penalization of higher entropies with lower variance that leads to lower average word topic entropies. The plot therefore indicates that the regularization does indeed result in lower entropies in words' latent topic distributions.

Next, we study the effect of regularization on document perplexity (shown in Figure 3.4(b)).

[2]The collapsed Gibbs sampler burns in after relatively few iterations since the reviews datasets that we use have only an average of 185 words per document.



(a) Average Word Distribution Entropy (on 10% heldout tuning set)



(b) Perplexity (on 10% heldout tuning set)

Figure 3.4: Studying the effect of Word Distribution Regularization.

Perplexity is a function of likelihood of the data and is defined as

$$2^{-\frac{\sum_{d=1}^{D}\sum_{i=1}^{N_d} log_2 p(w_{d,i})}{\sum_d N_d}} \tag{3.11}$$

It can be observed that many of the plots exhibit a "U-shape": i.e., the perplexity values dip

Kitchen	**Lowest:** 1.291: ice pillow cream pillows garlic sleep peeler great scoop firm knife sugar best butter neck support gears peeling love comfortable koolaid freezer bucket shaved juice buy head works perfect easily maker sweet soft nice better
	Highest: -1.292: coffee water machine time problem product don't works never bought doesn't unit replacement bought shipping send warranty didn't told quality light weeks working broken worked broke machine works bad arrived
DVD	**Highest:** 1.353: la de et assasination les pour garrison ville une est jet cats ce du sont archie qui il vie docteur dans edith pansori unanswered grande galactica cho tout beaucoup un li's que au
	Lowest: -1.309: movie film story bad don't people time think horror characters movies films plot scene acting scenes character little real end didn't never love watch watching better pretty things action thing seen girl actually
Electronics	**Highest:** 0.923: sound tv great quality headphones price speakers picture ear better set bass bought product don't look hd best sony little noise ears buy cable excellent lcd samsung right monitor pair fit looks nice pretty money reviews high volume mount cord
	Lowest: -1.464: unit product back service time buy work new bought year customer amazon don't support gps months warranty problem never weeks return days worked item sent working money replacement great garmin received battery told day didn't purchase map software tried
Books	**Highest:** 1.655: bogle agnes brady gregory book blacks lucy read walter brady's glue talents frances colonel persian great blues mae tooth conner dot julia hermione bram harris bront yellow bloom catherine penguin canyon love dorothy
	Lowest: -2.541: burr salem oswald jfk kennedy bangkok rifle monastery lou himis notovitch's notovitch parris oswald's harvey assassination president fired conspiracy da sonchai report fa dallas lama doctors tattoo tippit parkland lee caffeine monopoly dat aaron barrel document patsy j.d shots snide lennon witnesses wound max interpreter kimmel official texas pm
Movies	**Highest:** 0.723: film, time, character, story, life, characters, love, world, performance, films, director, scene, little, own, people, role, family, makes, john, takes, wife, scenes, american, audience, pretty, real, picture, sense, music, father, doesn, woman, death, cast, day, feel, simply, mother
	Lowest: -0.412: movie, film, bad, plot, movies, action, don, re, funny, doesn, people, little, films, ve, scenes, actually, isn, guy, scene, original, characters, seen, course, hard, effects, fun, didn, comedy, ll, minutes, script, look, acting, watch, star, series, special, lot, director, watching

Table 3.2: Topics with most positive and negative coefficients

below the value of the unregularized model (except when $K = 5$ for the books, DVD and kitchen datasets) when the variance value is set to 0.5, but rise again when the regularization is further

tightened and set to 0.2. This indicates that there is a "sweet spot" for the regularization, which is $\sigma_{l_w}^2 = 0.5$ in the case of the datasets studied here. We hypothesize that this is due to the polysemy freedom afforded by the unregularized model being overly expressive, and low variance values such as 0.2 reducing the freedom to a level where it is insufficient to represent the inherent polysemy in the corpus. In general, we use the optimal value of the hyperparameter corresponding to the best perplexity value on a 10% held out tuning set. The value thus determined is used for the remainder of experiments after folding in the tuning set into the training set.

Table 3.2 shows sample topics from a SLDA model with word latent role entropy regularization. Since the SLDA model has an associated classifier which uses topic proportions as input, we can determine the topic that is deemed to be most informative for star-rating by looking at the classifier co-efficients. The table shows the topics that are associated with the most positive and most negative co-efficients for each of the datasets.

For another qualitative treatment of the effects of word topic distribution regularization, we compare the latent role distributions as obtained from regularized and unregularized SLDA models of illustrative sample words. For the regularized models, we use a $\sigma_{l_w}^2$ value of 0.5. Firstly, we examine the word *slicer* from the kitchen product reviews dataset. In the context of kitchen appliance reviews, this word is fairly unambiguous and refers to the slicers in juicer appliances. In the regularized model, the word has the maximum mass on a topic that represents juicers (the top words in the topic are juice, juicer, juicing, pulp, omega, spout, carrots, wheatgrass) with 0.85 of the mass lying on this topic. In the unregularized model, 0.61 of the latent role distribution mass lies on a similar topic that discusses juicers. Therefore we see that the regularization leads to a larger mass to be assigned to the "right" topic for the word. Next, we look at the word *fidelity* from the electronics dataset. In the regularized model, the top two topics for this word are related to "headphones" (since fidelity refers to sound quality) and "printers" (since Fidelity is a commercial printer manufacturer) with the respective masses being 0.32 and 0.28 respectively. In contrast, in the unregularized model, the "headphones" and "printers" topic get a mass of 0.22 and 0.20 and a topic related to "DVD drives" gets a mass of 0.19. It can be seen from this example that the regularized model favors more peaky distributions which place more of the mass on a relatively smaller number of relevant topics.

Dataset	books	DVD	kitchen	electronics	movies
Range	1 - 5	1 - 5	1 -5	1 - 5	0 - 1
No regularization	2.169	2.042	1.986	1.885	0.218
With regularization	**2.136***	**2.003***	**1.761***	**1.874***	**0.208***
SVM	1.643	1.773	1.506	1.732	0.158
SVM + LDA (unregularized) features	1.614*	1.757	1.373	**1.456***	0.145*
SVM + LDA (with regularization) features	**1.592***	**1.748***	**1.356**	**1.456***	**0.142***

(starred entries indicate statistically significant entries)

Table 3.3: Effect of Word Distribution Regularization on MSE in star-rating prediction (computed using 10-fold cross-validation)

Finally, we test the performance of the model in predicting the star-ratings of reviews. We evaluate the star-rating predictions of models using mean-squared error (MSE), which is the square of the difference between the true and predicted ratings. Table 3.3 shows the mean squared error of the star-rating predictions for the different datasets. The table compares the error rates of the regularized models to the error rates of a baseline unregularized model. The number of topics K is set to either 10 or 20 depending on the perplexity observed during cross-validation for each datset. For the regularized model, we use use a variance values of 0.5 for $\sigma_{l_w}^2$, which showed the best results in the perplexity plot. It can be seen from the table that using word topic regularization consistently significantly improves the MSE in star-prediction. On average this improvement in error rate is 3.98%. The values in bold indicate the best performing model for a dataset and values marked with a $*$ indicate statistically significant improvements over the baselines, which is the unregularized SLDA model for the SLDA results and the SVM model for the SVM-based experiments.

We also run experiments where we add the topic distributions as additional features to the original bag of words vector representation and train an SVM regression model. We therefore study the effect of the topic distributions obtained with and without regularization in such a setting. We first evaluate the SVM regression model with a linear kernel as a baseline model using 10-fold cross validation for the star-rating prediction task. We then add the topic proportion distributions of doc-

uments obtained from a LDA model as additional features to the SVM model. The slack parameter value was set at 0.05 based on cross-validation. It can be seen from the results that adding topic model features helps in improving SVM regression performance and adding regularization further improves performance in all the datasets.

The results on the star-rating prediction task show that the regularization not only shows better perplexity, i.e. it returns a better language model, but that it also returns topic proportion distributions (θ_d) that encode more accurate information about the star-rating assigned to reviews.

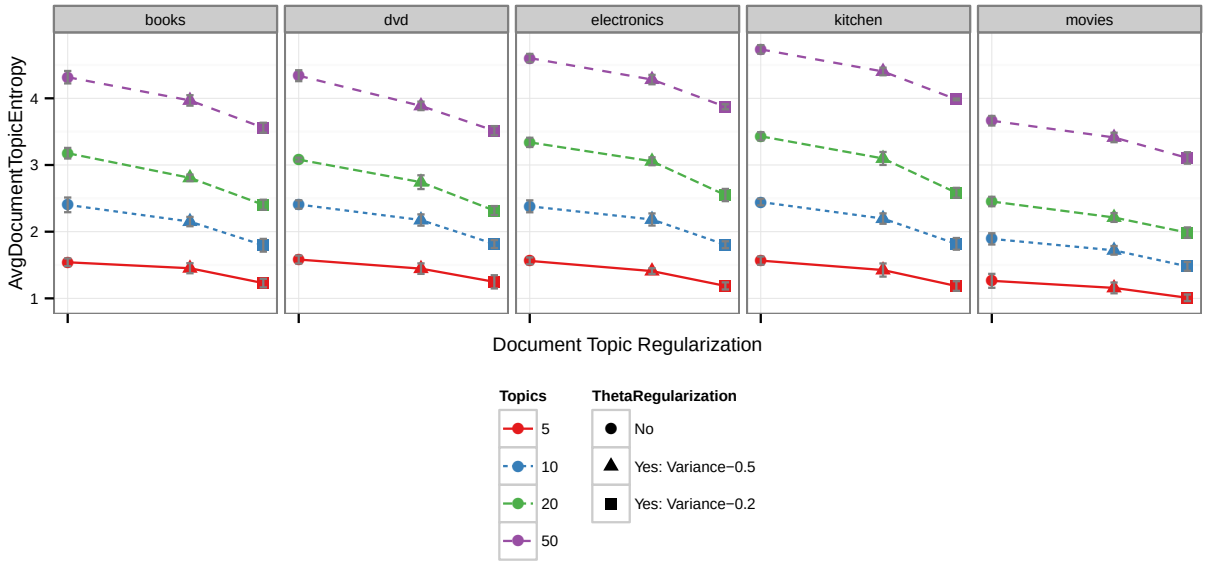
3.5.2 Effect of Document Topic Proportion Regularization

We also studied the effect of document topic proportion regularization on the entropy of $\overline{\mathbf{z}_d}$ averaged across all documents in the corpus. Figure **??** shows the plot of this value against different levels of regularization. The left points in each line indicate results with no regularization whereas the second and third points show results with $\sigma_{l_d}^2$ set to 0.5 and 0.2 respectively. We see that as the regularization is applied and more heavily enforced by lowering the variance, the average entropy is driven lower, indicating that the regularization has the intended effect.

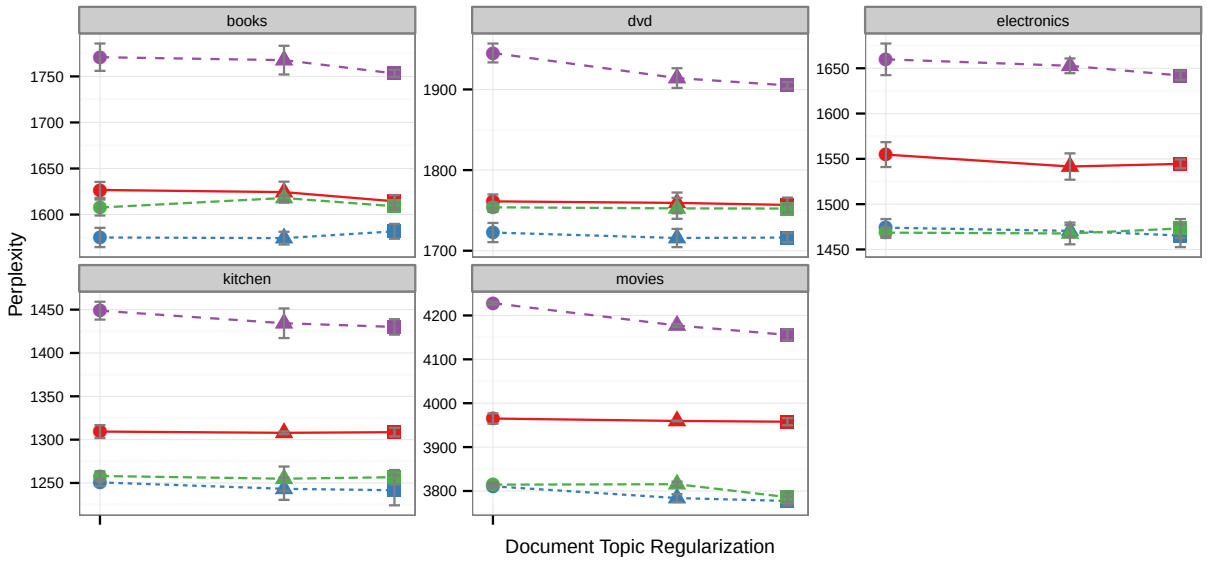
Next, we study the effect of topic proportion regularization on document perplexity in Figure 3.5(b). In all the datasets, the perplexity values decrease as the regularization is enforced and tightened. This effect is more prominently seen in the models with more topics. For instance, the purple lines at the top in each plot ($K = 50$), show a more noticeable drop in perplexity, whereas the solid red lines (with $K = 5$) show little movement with different levels of regularization. As the number of topics is increased, the unregularized model tends to use more diffuse distributions over topics, and the regularization helps in reducing this tendency, resulting in better perplexities.

From figures 3.4 and 3.5, we see that a variance value of 0.5 offers the best perplexity for all the datasets. In figure 3.6, we investigate the sensitivity of the hyperparameter value to external measures of performance, i.e., MSE in our experiments. From the plots, it can be seen that the MSE curves behave in a similar manner to the perplexity curves. This observation indicates that for this task, perplexity and MSE are correlated in terms of sensitivity to $\sigma_{l_w}^2$ and $\sigma_{l_d}^2$.

Figure 3.7 shows the relation between setting the hyperparameter of the Dirichlet prior to the document topic proportion distribution to control sparsity and using regularization. For these

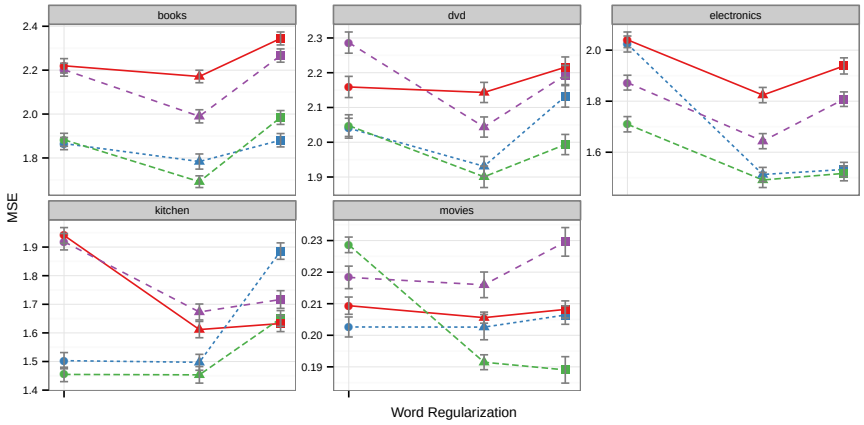


(a) Average Document Topic Proportion Distribution Entropy (on 10% heldout tuning set)

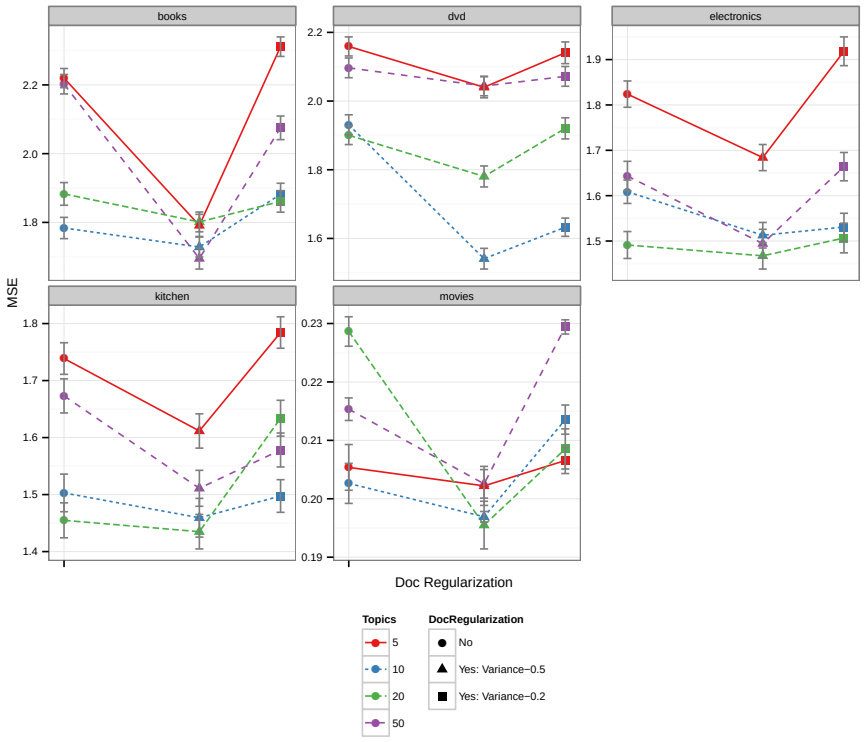


(b) Effect on Perplexity (on 10% heldout tuning set)

Figure 3.5: Effect of Document Topic Proportion regularization.



(a) Word Regularization



(b) Document Regularization

Figure 3.6: Effect of Regularization on MSE (computed using 10-fold cross-validation)

results, the number of topics was set to 10. The average entropy of the topic distributions for different values of the the symmetric Dirichlet hyperparameter (on the x-axis) and the regularization

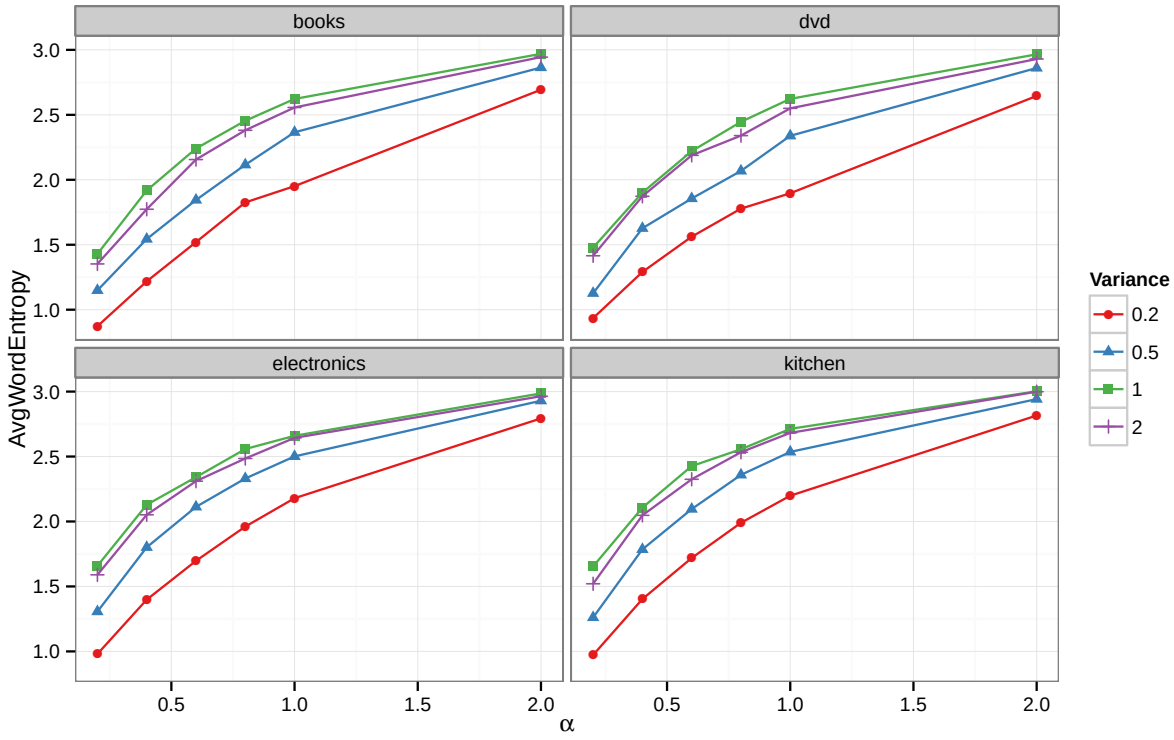


Figure 3.7: Comparison: Dirichlet hyperparameters vs. document topic proportion regularization (computed using 10-fold cross-validation)

variance value, i.e., $\sigma_{l_d}^2$ are plotted. It can be seen that for any constant α value, the average document topic distribution entropy decreases when regularization is used and when the variance value is decreased. It can also be seen that the entropy value decreases as the hyperparameter value is decreased. The lowest entropy value observed when regularization is used (i.e., when $\sigma_{l_d}^2$ is set to 0.2) can also be obtained by using a lower Dirichlet hyperparameter value. As discussed earlier, regularization of θ has the same effect as modifying the prior, which was not possible in the case of word latent role distribution regularization. In the latter case, modifying the parameters of the Dirichlet prior only affects the sparsity within a topic and does not directly address the distribution of a word across all topics.

Next, as with the previous set of experiments, we study the effect of regularization on star-prediction tasks, the results for which are shown in Table 3.4. For the results in the table, the variance hyperparameter for the regularization $\sigma_{l_d}^2$ was set to 0.5 (which is the optimal value based

Dataset	books	DVD	kitchen	electronics	movies
No regularization	2.169	2.042	1.986	1.885	0.218
With regularization	**2.121***	**2.031***	**1.784***	**1.743***	**0.210***
SVM	1.643	1.773	1.506	1.732	0.158
SVM + LDA (unregularized) features	1.614*	1.757	1.373	1.456	0.145*
SVM + LDA (with regularization) features	**1.601***	**1.750***	**1.364***	**1.454***	**0.143***

Table 3.4: Document Topic Regularization: Effect on MSE in star-rating prediction (computed using 10-fold cross-validation)

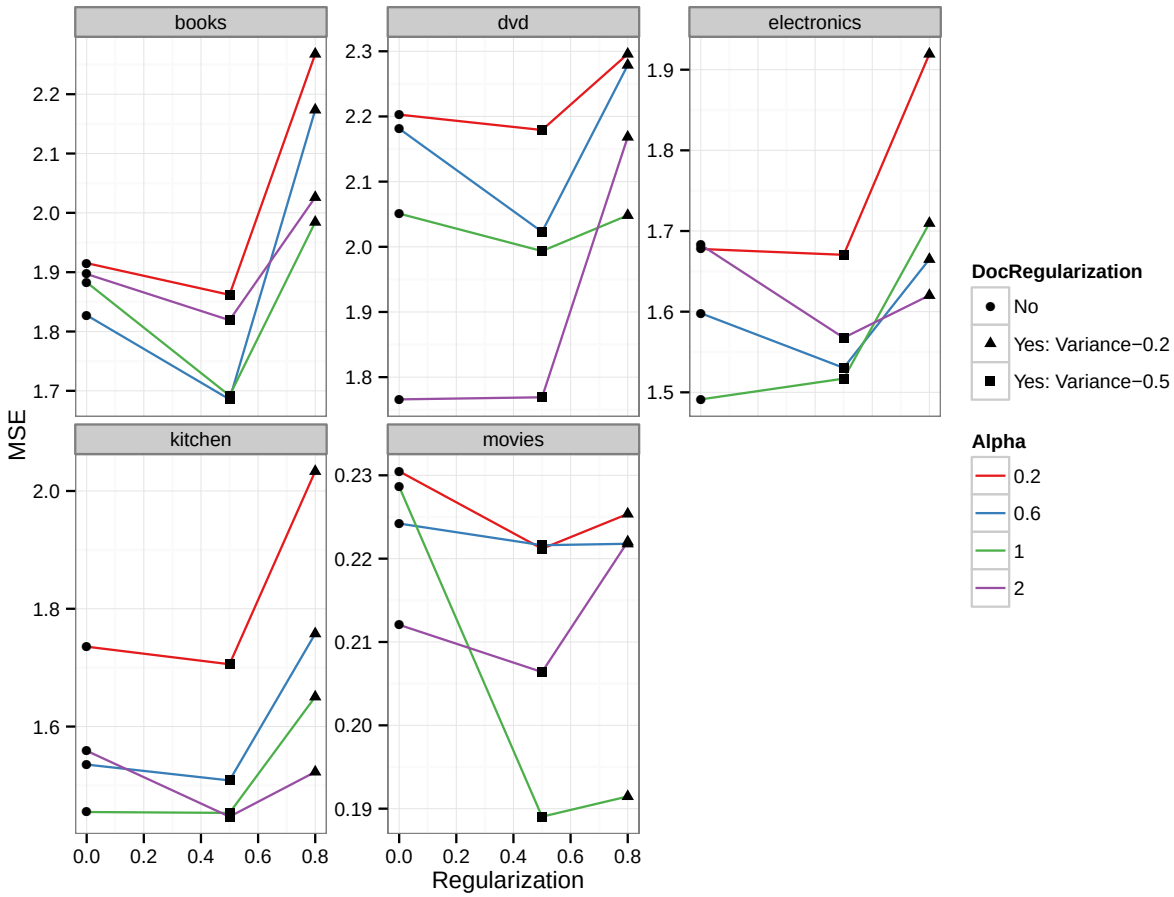


Figure 3.8: Varying Dirichlet hyperparameter and document topic proportion regularization (MSE computed using 10-fold cross-validation)

on the perplexity plots) and α to 1.0. It can be seen from the table that adding the regularization improves the MSE for all the datasets by 4.82% on average. The improvements for all the datasets are statistically significant at p=0.05 using the Wilcoxon paired sign test. These results suggest that more peaked topic distributions describe the corpus better and also contain more information about the star-rating.

Finally, we study the effect of modifying both the Dirichlet hyperparamater and the regularization hyperparameter $\sigma_{l_d}^2$ on the MSE in predicting star-ratings. In figure 3.8, we see that just as in regularization, there is a sweet spot for the Dirichlet hyperparameter value which is either 0.6 or 1.0 in the datasets studied here. In 4 out of the 5 datasets, we see that the best MSE value is observed with $\alpha = 1$ and with $\sigma_{l_d}^2$ set to 0.5, which indicates that modifying the Dirichlet parameter does not translate to performance gains in MSE, although we see that we can get the same average word entropy characteristics as achieved by regularization (figure 3.7). For non-optimal α values, the gains introduced by regularization can also be obtained by moving to a better α value. The sparsity introduced by a low Dirichlet hyperparameter and by lowering the entropy of θ lead to different kinds of sparsity, and empirically we see that the topic proportion obtained after fitting a regularized model, has better predictive power. Therefore, while sparsity can be introduced by altering the Dirichlet hyperparameter, entropy reduction in θ using the regularization framework is advantageous in terms of star-rating prediction on these datasets. We intend to investigate the relationship between the Dirichlet and variance hyperparameters in detail in future work.

3.6 Related Work

The general idea of imposing constraints in the inferred posterior distributions has been addressed in different ways previously by researchers. Ganchev et al. [2010] proposed Posterior Regularization (PR), a method to incorporate indirect supervision via constraints on posterior distributions of probabilistic models with latent variables. They demonstrate the use of the technique in models for several tasks such as POS induction and word alignment. While the approach proposed in this chapter is similar in spirit to PR, in that both approaches provide a method for preferences for the posteriors of latent variables to be specified, there are significant differences in the manner in which

the constraints are imposed. The PR framework works by decoupling the constraint requirements from the modeling process. An EM procedure is proposed which alternates between training the unconstrained model and ensuring that the constraints are satisfied. In contrast, the regularization approach proposed here introduces the regularization terms directly into the graphical model, in a manner which only minimally affects the approximate inference procedure (by the addition of a few terms), through the use of pseudo-observed noisy copy of aggregate function values.

Chang et al. [2007] introduced Constraint Driven Learning (CODL) where domain knowledge is introduced into semisupervised learning models. Similar to the approach used in PR, the objective function of the model being trained is modified to add a penalty which is proportional to the distance between desired values of the inferred distributions and the actual distribution. An EM approach is used to train models using the modified objective function which alternates between fitting the original objective function and satisfying constraints.

Liang et al. [2009] presented a new framework for introducing domain knowledge based constraints in a Bayesian setting called *measurements*. Their approach and ours share some characteristics — the objective function has a prior distribution, the likelihood term and a noise model which penalize deviations from desired values of functions over inferred variables. Liang et al. [2009] use a box-based noise model, in contrast to the Gaussian based model that is used in this chapter. Their work focuses on using the framework for tackling structured learning tasks and they propose different ways to fit the model to data using approximation techniques like mean field factorization. Here, we focus exclusively on latent variable mixed-membership models, where the addition of the regularization does not require major changes to the approximate inference procedure.

Wang and Blei [2009] presented sparseTM, a nonparametric topic model in which sparsity and smoothness are controlled separately. A bank of Bernoulli variables is used to determine which terms are allowed to be part of a topic which causes each topic to be sparse. This form of sparsity is in contrast to the sparsity introduced by word latent role distribution which induces sparsity of a word *across* topics. However, it is likely that strenghtening the sparsity in sparseTM will cause the word latent role distribution τ_w to have low entropy as well.

Eisenstein et al. [2011] introduced SAGE, a sparse additive generative model in which topics are modeled as deviations from a central background topic. Sparsity is introduced by limiting the

number of terms whose probabilities can deviate from the background topic. Like in the case of sparseTM, SAGE introduces sparsity within the topic, but in a different manner.

Mann and McCallum [2010] also proposed a general framework to introduce preferences in model expectations by adding terms called *generalized expectation (GE) criteria* to the objective function. Examples of such criteria were explored in the domain of log-linear models. The approach in this chapter shares similarities to the GE framework in that the regularization operates on entropies of distributions of inferred latent variables; moreover, the GE approach specifies constraints in the form of expected values over the inferred distributions, which are similar to the pseudo-observed values used in this work. However, the manner in which deviations from expectations are penalized in our approach is different from the criteria used in Mann and McCallum [2010]: the method introduced in this chapter proposes that a desired value is drawn from a distribution parameterized by the inferred latent variables' values, whereas the GE framework uses KL divergence to penalize the deviation from an expected distribution. The GE framework has not been applied to latent variable mixed-membership models, as far as we know.

Newman et al. [2011] presented a method to regularize topic models to produce coherent topics. In this approach, a pre-computed matrix of word-similarities from external data (Wikipedia) is used to construct a prior for the topic distributions. This regularization approach differs from the framework used in this paper in that it is aimed at producing topics that respect external word similarities. This is in contrast to our approach that is designed to control the latent structure properties without using external data.

Regularization by entropy has been used previously for semi-supervised learning in Grandvalet and Bengio [2005], Jiao et al. [2006] and Corduneanu and Jaakkola [2005] where entropy based regularizers are used to constrain the unknown labels of unlabeled data points. Celeux and Soromenho [1996] also use criteria based on entropy to determine the optimal number of clusters in mixture models. The approach presented here uses entropy for a different purpose, i.e., to impose preferences on the mixed-membershipness of words/entities. Regularization in models based on LDA have also been previously proposed in works such as Cai et al. [2008] and Mei et al. [2008], which use a regularization term in the likelihood expression to remove the independence assumption between documents by placing them on a manifold.

Chang and Blei [2010] and Gruber et al. [2008] presented models that jointly model documents and the network between them. The joint models encourages topics to have regularity in order to explain a network of documents, based on characteristics of documents that are linked together, rather than characteristics of the observed and latent variables as presented here.

Opinion mining and sentiment analysis has become an active area of research in the last decade [Pang and Lee, 2008]. The aim in this line of work is to analyze the sentiments expressed in online communities via reviews on movies, products, hotels etc. Often the task is simplified by framing the problem as extracting the opinion polarities or numeric ratings (such as star ratings on amazon.com) [Pang and Lee, 2005]. Qu et al. [2010] tackled the problem by replacing the bag-of-words representation of review text with a bag-of-opinions representation that uses linguistic cues to get features that contain more signal about the opinion expressed. Popescu and Etzioni [2005] also address the problem or feature representation and present a system that represents revies as a set of feature and opinion tuples. Titov and McDonald [2008a] used topic models to extract aspects ratings [Hu and Liu, 2004] from reviews. Lerman et al. [2009] provides an overview of different sentiment summarizers which look at reviews and provide a short summary about the sentiment expressed in it.

3.7 Conclusion

In this chapter, we presented an regularization approach to obtain finer control over the latent structure obtained from mixed-membership latent variable models. We used the method to sparsify topic models; firstly we used it to softly constrain words' ability to participate in multiple topics thus providing a way to control the ability of the model to permit polysemy. We then used the entropic regularization approach to make the topic proportion distribution of documents sparse thus permitting LDA-like models to span the spectrum from a mixture of multinomials to a fully unconstrained LDA model. Our experiments show that the word entropy and document topic regularization result in better perplexity and mean-squared error scores in the star-rating prediction task because it enables the model to utilize sentiment-indicative words more efficiently. The work described in this chapter was published earlier in Balasubramanyan and Cohen [2013].

Chapter 4

Bayesian Formulation

4.1 Motivation

In chapter 3 (Figure 3.2), we presented a method to regularize latent variable models. Specifically, we use the proposed framework to control the entropy of words' latent role distribution entropies. The method therefore provided additional modeling flexibility than is normally present in such models. From a purely Bayesian standpoint, the natural way to introduce such preferences is to use suitable prior distributions that encode the preferences we wish to impose. In the case of words' latent role distribution properties, this introduces complications since the distributions are not explicitly sampled in the generative story. They are rather a byproduct obtained from distributions that are generated. Specifically, words' latent role distributions are obtained by observing a word's probability in each topic multinomial that is drawn iid from a Dirichlet prior. To use a purely Bayesian approach would require the Dirichlet prior distributions to be replaced with a different distribution that serves as a prior for the set of topic multinomials. Since we desire to introduce dependencies between a word's probability in each of the topics, the new distribution necessarily requires that the topic multinomials are no longer independent of each other. Furthermore, it can be easily seen that replacing the Dirichlet prior distribution causes a loss of conjugacy that is enjoyed by the Dirichlet prior and multinomial topic distribution pair. The conjugacy afforded when using a Dirichlet prior makes it simple to evaluate integrals needed in the collapsed Gibbs sampler. Modifying the prior will therefore negate this computational convenience. When it is impossible to

obtain a closed form expression for the likelihood integral, alternate inference schemes will need to be employed since the use of collapsed Gibbs sampling, which requires the computation of the conditional likelihood, will not be possible. In this chapter, we present a new prior distribution for topic multinomials that permits us to specify relations of word's probabilty across hitherto independent multinomials. The use of the new prior comes at the cost of more complex inference using a Metropolis-Hastings sampling scheme. This new prior captures the same intuition as the word-entropy regularization introduced in Section 3.2 but implements that intuition differently. We will also show that it has a similar effect experimentally, in that samples generated from models using this prior with the computationally expensive Metropolis-Hastings algorithm, largely overlap with samples generated from the model of Section 3.2

4.2 Entropically Constrained LDA

In this model we start with LDA and and define a new prior distribution for the topic multinomials $\beta_k, k \in \{1, \ldots, K\}$. In the LDA model, the topic multinomials are drawn from a Dirichlet prior i.e. the density for a multinomial β_k is given by

$$\mathsf{Dir}(\beta_k|\gamma) = \frac{1}{\Delta(\gamma)} \prod_v \beta_{k,v}^{\gamma_v - 1} \tag{4.1}$$

Here $\Delta(\gamma)$ is the normalizing constant and is defined as $\frac{\prod_v \Gamma(\gamma_v)}{\Gamma(\sum_v \gamma_v)}$. When the hyperparameters $\gamma_v,\ v \in 1, \ldots, V = \gamma$ are equal in value, this is termed a symmetric Dirichlet prior.

Our goal when introducing the regularization of the latent role distribution of words (Section 3.2) was to introduce dependencies between $\beta_{k,v}$. This permitted us among other things to specify that a word's latent role distribution should have low entropy. To introduce the desired dependencies via a prior rather than by introducing regularization terms, we propose a new prior distribution over the set of K multinomials. The newly proposed *Entropically Coupled Dirichlet (ECD)* distribution over $\beta_k, k \in 1, \ldots, K$ is defined as:

$$\mathsf{ECD}(\boldsymbol{\beta}|\gamma, \sigma_{l_v}^2) = \frac{1}{C} \left[\prod_k \prod_v \beta_{k,v}^{\gamma - 1} \right] \left[\prod_v \exp\left(\frac{-\left(-\sum_k \beta'_{k,v} \log \beta'_{k,v}\right)^2}{2\sigma_{l_v}^2} \right) \right]; \beta'_{k,v} = \frac{\beta_{k,v}}{\sum_{k'} \beta_{k',v}} \tag{4.2}$$

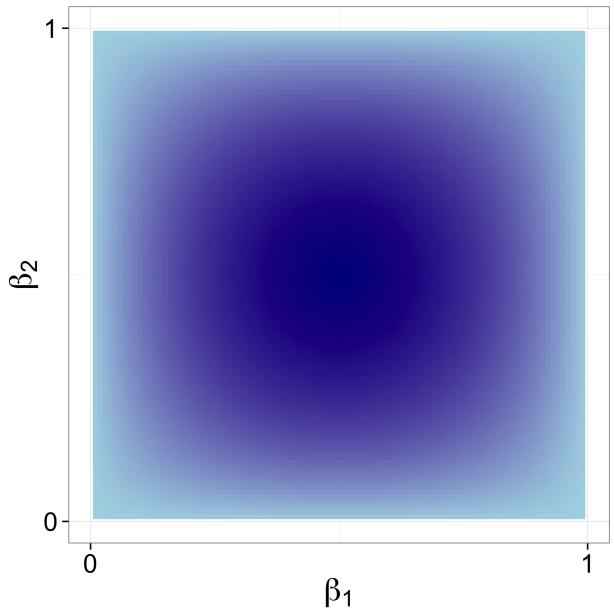
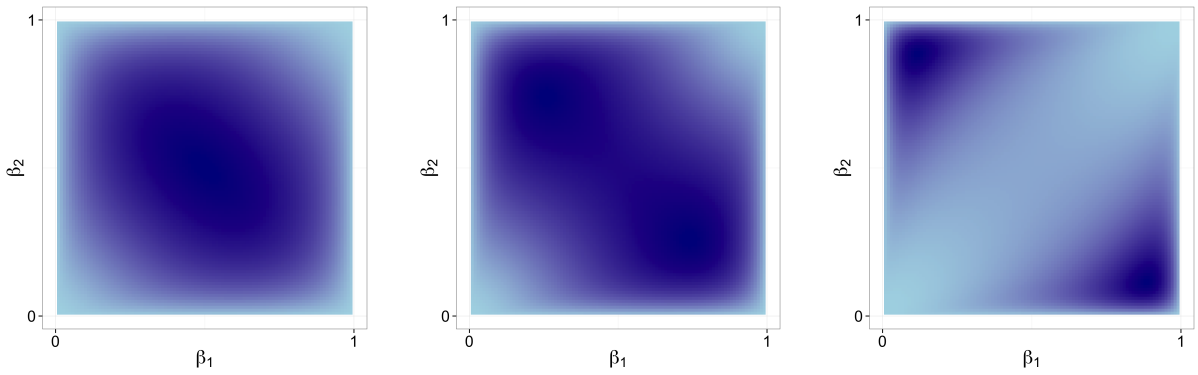


Figure 4.1: Density of Two Binomials with Beta prior

To illustrate the change introduced when using a ECD prior in lieu of Dirichlet priors, let us consider a toy example. We consider a two topic model where the vocabulary size is 2. Therefore each of the two topics is a binomial rather than a multinomial. Figure 4.1 shows the density of drawing 2 binomials from a Beta prior (which is the analog of a Dirichlet prior for binomials) with hyperparameters $2, 2$. Topic 1, i.e., β_1 is represented on the x-axis with the x-axis values indicating the probability of the first word in topic 1. Similarly, topic 2 i.e. β_2 is represented in the y-axis. Darker points in the plot indicate a higher density. It can be seen that the distribution is symmetric around the diagonal and anti-diagonal and the mass is heaviest at $0.5, 0.5$ where both topics have equal probabilities for the two words in the vocabulary.

Figure 4.2 shows the density of drawing the same two topics β_1 and β_2 using an ECD prior ($\gamma = 2$). The three subfigures show the densities with different values for the hyperparameter $\sigma_{l_v}^2$. In these plots, it can be seen that the probability mass is no longer symmetric around the diagonal, and that the probability mass is skewed towards the top left and bottom right corners. These two corners represent low entropy topic distributions. For instance, the bottom right corner represents a point where $\beta_1 = \{1, 0, 0.0\}$ and $\beta_2 = \{0.0, 1.0\}$. At this point, both words in the vocabulary have a Shannon entropy of 0. We further see that as the hyperparameter value is increased, this

effect is more pronounced. Therefore, we see that using an ECD prior biases the model towards drawing a set of topics where words have low latent role distributions.



(a) $\sigma_{l_v}^2 = 1$ (b) $\sigma_{l_n}^2 = 0.5$ (c) $\sigma_{l_n}^2 = 0.2$

Figure 4.2: Density of Two Binomials with an ECD prior

4.2.1 Inference in LDA with ECD Priors

The generative story behind LDA with an ECD prior, of a corpus with M documents is fully described as follows:

$$
\begin{aligned}
&\boldsymbol{\beta} \sim \text{ECD}(\gamma, \sigma_{l_v}^2), \quad \text{where } \boldsymbol{\beta} = \langle \beta_1, \ldots, \beta_K \rangle \text{ i.e. a vector of } K \ V\text{-dimensional multinomials} \\
&\theta_m \sim \text{Dir}(\alpha), m \in 1, \ldots, M \\
&z_{m,n} \sim \text{Mult}(\theta_m), n \in 1, \ldots, N_m \\
&w_{m,n} \sim \text{Mult}(\beta_{z_{m,n}})
\end{aligned} \tag{4.3}
$$

The likelihood of the data is given by:

$$
p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}, \boldsymbol{\beta} | \alpha, \gamma, \sigma_{l_v}^2) = \left[\prod_{m=1}^{M} p(\theta_m | \alpha) \prod_{n=1}^{N_m} p(z_{m,n} | \theta_m) p(w_{m,n} | \beta_{z_{m,n}}) \right] \text{ECD}(\boldsymbol{\beta} | \gamma, \sigma_{l_v}^2) \tag{4.4}
$$

In previous chapters, we employed collapsed Gibbs sampling to perform approximate inference. Here we will show that collapsed Gibbs sampling is no longer possible when the Dirichlet priors for $\boldsymbol{\beta}$ are replaced with the ECD distribution described above. Therefore we propose an alternate MCMC sampling scheme based on the Metropolis-Hastings algorithm. It can of course be noted

that while collapsed Gibbs sampling is not possible, a Gibbs sampling scheme that samples not just $\mathbf{z}$ variables, but also values for $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ is still possible. However, we anticipate that the process will be inefficient due to the high dimensionality of the variables to be sampled.

We now describe the proposed Metropolis-Hastings inference scheme. To begin, we define the following terms, which are dependent only on $\mathbf{w}$ and $\mathbf{z}$ for convenience.

$$n_{mk} = \sum_{n=1}^{N_m} \delta(z_{m,n} = k), \quad \overrightarrow{n_m} = \langle n_{m1}, n_{m2}, \ldots, n_{mK} \rangle \tag{4.5}$$
$$n_{kv} = \sum_{m=1}^{M} \sum_{n=1}^{N_m} \delta(w_{m,n} = v) \times \delta(z_{m,n} = k), \quad n_k = \sum_v n_{kv}$$

The likelihood (Equation 4.4) can be re-expressed using these terms as

$$\begin{aligned} p(\boldsymbol{\theta}, \mathbf{z}, \mathbf{w}, \boldsymbol{\beta} | \alpha, \gamma, \sigma_{l_v}^2) = & \left[\prod_{m=1}^{M} \frac{1}{\Delta(\gamma)} \prod_k \theta_{m,k}^{\gamma-1} \prod_k \theta_{m,k}^{n_{mk}} \right] \left[\prod_k \prod_v \beta_{k,v}^{n_{kv}} \right] \text{ECD}(\boldsymbol{\beta} | \gamma, \sigma_{l_v}^2) \\ = & \left[\prod_{m=1}^{M} \frac{1}{\Delta(\gamma)} \prod_k \theta_{m,k}^{\gamma+n_{mk}-1} \right] \left[\prod_k \prod_v \beta_{k,v}^{n_{kv}} \right] \text{ECD}(\boldsymbol{\beta} | \gamma, \sigma_{l_v}^2) \end{aligned} \tag{4.6}$$

Integrating out $\boldsymbol{\theta}$ using $\int \prod_k \theta_{m,k}^{\alpha+n_{mk}-1} d\theta_m = \Delta(\overrightarrow{n_m} + \alpha)$ and re-arranging terms, we get

$$p(\mathbf{z}, \mathbf{w}, \boldsymbol{\beta} | \alpha, \gamma, \sigma_{l_v}^2) = \left[\prod_{m=1}^{M} \frac{\Delta(\overrightarrow{n_m} + \alpha)}{\Delta(\alpha)} \right] \left[\prod_k \prod_v \beta_{k,v}^{n_{kv}} \right] \text{ECD}(\boldsymbol{\beta} | \gamma, \sigma_{l_v}^2) \tag{4.7}$$

It can now be seen that integrating out $\boldsymbol{\beta}$ is not possible since ECD is not conjugate to multinomials, and the integral has no closed form solution.

Since

$$p(\mathbf{z}, \boldsymbol{\beta} | \mathbf{w}, \alpha, \gamma, \sigma_{l_v}^2) = \frac{p(\mathbf{z}, \mathbf{w}, \boldsymbol{\beta} | \alpha, \gamma, \sigma_{l_v}^2)}{p(\mathbf{w} | \alpha, \gamma, \sigma_{l_v}^2)} \text{ and}$$
$$p(\mathbf{w} | \alpha, \gamma, \sigma_{l_v}^2) = \int \int \left[\prod_{m=1}^{M} \frac{\Delta(\overrightarrow{n_m} + \alpha)}{\Delta(\alpha)} \right] \left[\prod_k \prod_v \beta_{k,v}^{n_{kv}} \right] \text{ECD}(\boldsymbol{\beta} | \gamma, \sigma_{l_v}^2) \, d\boldsymbol{\beta} \, d\mathbf{z} = F(\alpha, \gamma, \sigma_{l_v}^2)$$

it can be seen that

$$p(\mathbf{z}, \boldsymbol{\beta} | \mathbf{w}, \alpha, \gamma, \sigma_{l_v}^2) \propto \left[\prod_{m=1}^{M} \frac{\Delta(\overrightarrow{n_m} + \alpha)}{\Delta(\alpha)} \right] \left[\prod_k \prod_v \beta_{k,v}^{n_{kv}} \right] \text{ECD}(\boldsymbol{\beta} | \gamma, \sigma_{l_v}^2) \tag{4.8}$$

To sample $\mathbf{z}$ and $\boldsymbol{\beta}$ from this distribution from the above distribution, we propose a Metropolis-Hastings based algorithm.

Metropolis Hastings Algorithm

The Metropolis-Hastings (MH) algorithm is a Markov Chain Monte Carlo scheme to obtain samples from a distribution in cases where it is not possible to directly sample from the desired distribution. Possible reasons for the inability to directly sample includes the lack of a closed form and high computational expense. Let $p(\mathbf{\Theta}|y)$ be the distribution we want a sample from. The MH algorithm proceeds as follows

- $\mathbf{\Theta}^{(0)} \leftarrow x$
- for i $= 0$ to T
  - draw $\widetilde{\mathbf{\Theta}} \sim q(\mathbf{\Theta}|\mathbf{\Theta}^{(i)})$
  - set $\mathbf{\Theta}^{(i+1)} \leftarrow \widetilde{\mathbf{\Theta}}$ with probability $a(\mathbf{\Theta}^{(i)}, \widetilde{\mathbf{\Theta}})$
  - else set $\mathbf{\Theta}^{(i+1)}$ to $\mathbf{\Theta}^{(i)}$

$$a(c,d) = \min\left\{1, \frac{p(d|y)q(c|d)}{p(c|y)q(d|c)}\right\} \tag{4.9}$$

The algorithm looks like a stochastic hill climbing algorithm, and uses $q(\mathbf{\Theta}|\mathbf{\Theta}^{(i)})$, which is called the *proposal distribution*, to propose moves in state space. Since only ratios of probabilities of proposed states need to be computed, the need to compute normalizing constants (which are often expensive) is removed. The acceptance ratio $a(c,d)$ indicates if the newly proposed state returned by q leads to a better state. It accounts for the fact that the proposal density is not the target density p.

We use a proposal distribution — $q(\widetilde{\mathbf{z}}, \widetilde{\boldsymbol{\beta}}|\mathbf{z}, \boldsymbol{\beta})$ for the model proposed in 4.3 that follows the collapsed Gibbs sampler used for the regularized LDA model. Each call to q results in an update for a single $z_{m,n}$, the new value for which is sampled using:

$$q(z_{m,n} = k|\alpha, \gamma, \sigma_{l_v}^2, \mathbf{w}, \mathbf{z}^{\neg m,n}) \propto \quad (n_{mk}^{\neg m,n} + \alpha)\frac{n_{kw_{m,n}}^{\neg m,n} + \gamma}{\sum_{v'} n_{kv'}^{\neg m,n} + |V|\gamma}\exp\left(\frac{-(-\sum_k \tau_{kw_{m,n}} \log\ \tau_{kw_{m,n}})^2}{2\sigma_{l_v}^2}\right) \tag{4.10}$$

$\tau_{kw_{m,n}}$ is defined as $\frac{n_{kw_{m,n}}}{\sum_{k'} n_{k'w_{m,n}}}$ and $\widetilde{\boldsymbol{\beta}}$ is set to its MLE estimate derived from the updated $\widetilde{\mathbf{z}}$.

Note that $\widetilde{\mathbf{z}}$ differs from $\mathbf{z}$ in that $z_{m,n}$ has changed from k_{old} to k_{new}. We can now compute the counts $\widetilde{n}$ using the new sample $\widetilde{\mathbf{z}}$ and update the affected values in $\widetilde{\boldsymbol{\beta}}$ i.e. $\widetilde{\beta}_{k_{old}}$ and $\widetilde{\beta}_{k_{new}}$ using

$$\widetilde{\beta}_{k_{old},v} = \frac{\widetilde{n}_{k_{old}v} + \gamma}{\widetilde{n}_{k_{old}} + V\gamma}, \quad \widetilde{\beta}_{k_{new},v} = \frac{\widetilde{n}_{k_{new}v} + \gamma}{\widetilde{n}_{k_{new}} + V\gamma} \quad v \in 1, \ldots, V \tag{4.11}$$

Further it can be seen that $\widetilde{n}$ is the same as n except for

$$\widetilde{n}_{mk_{new}} = n_{mk_{new}} + 1, \qquad \widetilde{n}_{mk_{old}} = n_{mk_{old}} - 1$$
$$\widetilde{n}_{k_{new}w_{m,n}} = n_{k_{new}w_{m,n}} + 1, \qquad \widetilde{n}_{k_{old}w_{m,n}} = n_{k_{old}w_{m,n}} - 1$$

The new values $\widetilde{\mathbf{z}}$ and $\widetilde{\boldsymbol{\beta}}$ are accepted with the probability $\min\left(1, \frac{p(\widetilde{\mathbf{z}},\widetilde{\boldsymbol{\beta}}|\mathbf{w},\alpha,\gamma,\sigma_{l_v}^2)q(\mathbf{z},\boldsymbol{\beta}|\widetilde{\mathbf{z}},\widetilde{\boldsymbol{\beta}})}{p(\mathbf{z},\boldsymbol{\beta}|\mathbf{w},\alpha,\gamma,\sigma_{l_v}^2)q(\widetilde{\mathbf{z}},\widetilde{\boldsymbol{\beta}}|\mathbf{z},\boldsymbol{\beta})}\right)$.
The second term in the min expression above can be expanded as

$$\underbrace{\frac{\Delta(\overrightarrow{\widetilde{n}_m} + \alpha)}{\Delta(\overrightarrow{n_m} + \alpha)}}_{\text{Part 1}} \underbrace{\frac{\prod_v \widetilde{\beta}_{k_{old},v}^{\widetilde{n}_{k_{old}v}} \widetilde{\beta}_{k_{new},v}^{\widetilde{n}_{k_{new}v}}}{\prod_v \beta_{k_{old},v}^{n_{k_{old}v}} \beta_{k_{new},v}^{n_{k_{new}v}}}}_{\text{Part 2}} \underbrace{\frac{\mathsf{ECD}(\widetilde{\boldsymbol{\beta}}|\gamma,\sigma_{l_v}^2)}{\mathsf{ECD}(\boldsymbol{\beta}|\gamma,\sigma_{l_v}^2)}}_{\text{Part 3}} \times \tag{4.12}$$
$$\underbrace{\frac{(n_{mk_{old}} - 1 + \alpha)^{\frac{n_{k_{old}w_{m,n}} - 1 + \gamma}{n_{k_{old}} - 1 + V\gamma}}}{(n_{mk_{new}} + \alpha)^{\frac{n_{k_{new}w_{m,n}} + \gamma}{n_{k_{new}} + V\gamma}}}}_{\text{Part 4}} \times \underbrace{\exp\left(\frac{-(-\sum_k \tau_{kw_{m,n}} \log\ \tau_{kw_{m,n}})^2 + (-\sum_k \widetilde{t}_{kw_{m,n}} \log\ \widetilde{t}_{kw_{m,n}})^2}{2\sigma_{l_v}^2}\right)}_{\text{Part 5}}$$

where $\widetilde{t}_{kw_{m,n}}$ is defined as $\frac{\widetilde{n}_{kw_{m,n}}}{\sum_{k'} \widetilde{n}_{k'w_{m,n}}}$. It can be seen that the normalizing constants for $q(\widetilde{\mathbf{z}}, \widetilde{\boldsymbol{\beta}}|\mathbf{z}, \boldsymbol{\beta})$ and $q(\mathbf{z}, \boldsymbol{\beta}|\widetilde{\mathbf{z}}, \widetilde{\boldsymbol{\beta}})$ are equal and therefore cancel out.

Using the relation $\Gamma(n) = (n-1)\Gamma(n-1)$, Part 1 from Equation 4.12 can be reduced as follows.

$$\begin{aligned}
&\frac{\prod_k \Gamma(\widetilde{n}_{mk} + \alpha)}{\Gamma(\sum_k (\widetilde{n}_{mk} + \alpha))} \frac{\Gamma(\sum_k (n_{mk} + \alpha))}{\prod_k \Gamma(n_{mk} + \alpha)} \\
&= \frac{\Gamma(\widetilde{n}_{mk_{old}} + \alpha)\Gamma(\widetilde{n}_{mk_{new}} + \alpha)}{\Gamma(n_{mk_{old}} + \alpha)\Gamma(n_{mk_{new}} + \alpha)} \\
&= \frac{\Gamma(n_{mk_{old}} - 1 + \alpha)\Gamma(n_{mk_{new}} + 1 + \alpha)}{\Gamma(n_{mk_{old}} + \alpha)\Gamma(n_{mk_{new}} + \alpha)} \\
&= \frac{n_{mk_{new}} + \alpha}{n_{mk_{old}} - 1 + \alpha}
\end{aligned} \tag{4.13}$$

Part 2 of Equation 4.12 can be expanded by using point estimates for $\boldsymbol{\beta}$ as:

$$\frac{\left(\frac{n_{k_{old}w_{m,n}}-1+\gamma}{n_{k_{old}}-1+V\gamma}\right)^{(n_{k_{old}w_{m,n}}-1+\gamma-1)}}{\left(\frac{n_{k_{old}w_{m,n}}+\gamma}{n_{k_{old}}+V\gamma}\right)^{n_{k_{old}w_{m,n}}+\gamma-1}}\frac{\left(\frac{n_{k_{new}w_{m,n}}+1+\gamma}{n_{k_{new}}+1+V\gamma}\right)^{(n_{k_{new}w_{m,n}}+1+\gamma-1)}}{\left(\frac{n_{k_{new}w_{m,n}}+\gamma}{n_{k_{new}}+V\gamma}\right)^{n_{k_{new}w_{m,n}}+\gamma-1}}\times$$
$$\prod_{v\neq w_{m,n}}\frac{\left(\frac{n_{k_{old}v}+\gamma}{n_{k_{old}}-1+V\gamma}\right)^{n_{k_{old}v}+\gamma-1}}{\left(\frac{n_{k_{old}v}+\gamma}{n_{k_{old}}+V\gamma}\right)^{n_{k_{old}v}+\gamma-1}}\frac{\left(\frac{n_{k_{new}v}+\gamma}{n_{k_{new}}+1+V\gamma}\right)^{n_{k_{new}v}+\gamma-1}}{\left(\frac{n_{k_{new}v}+\gamma}{n_{k_{new}}+V\gamma}\right)^{n_{k_{new}v}+\gamma-1}}$$

Part 3 is re-written as:

$$\prod_{v}\exp\left(\frac{-\left(-\sum_k\widetilde{\beta}'_{k,v}\log\widetilde{\beta}'_{k,v}\right)^2+\left(-\sum_k\beta'_{k,v}\log\beta'_{k,v}\right)^2}{2\sigma_{l_v}^2}\right)$$

After merging Part 1 with Part 4, the sample from the proposal distribution (Equation 4.12) is therefore accepted with a probability of

$$\begin{aligned}
&\min\left(1,\frac{(n_{mk_{old}}-1+\alpha)}{(n_{mk_{new}}+\alpha)}\frac{\left(\frac{n_{k_{old}w_{m,n}}-1+\gamma}{n_{k_{old}}-1+V\gamma}\right)^{n_{k_{old}w_{m,n}}+\gamma-1}}{\left(\frac{n_{k_{old}w_{m,n}}+\gamma}{n_{k_{old}}+V\gamma}\right)^{n_{k_{old}w_{m,n}}+\gamma-1}}\frac{\left(\frac{n_{k_{new}w_{m,n}}+1+\gamma}{n_{k_{new}}+1+V\gamma}\right)^{(n_{k_{new}w_{m,n}}+\gamma)}}{\left(\frac{n_{k_{new}w_{m,n}}+\gamma}{n_{k_{new}}+V\gamma}\right)^{(n_{k_{new}w_{m,n}}+\gamma)}}\times\right.\\
&\prod_{v\neq w_{m,n}}\frac{\left(\frac{n_{k_{old}v}+\gamma}{n_{k_{old}}-1+V\gamma}\right)^{n_{k_{old}v}+\gamma-1}}{\left(\frac{n_{k_{old}v}+\gamma}{n_{k_{old}}+V\gamma}\right)^{n_{k_{old}v}+\gamma-1}}\frac{\left(\frac{n_{k_{new}v}+\gamma}{n_{k_{new}}+1+V\gamma}\right)^{n_{k_{new}v}+\gamma-1}}{\left(\frac{n_{k_{new}v}+\gamma}{n_{k_{new}}+V\gamma}\right)^{n_{k_{new}v}+\gamma-1}}\\
&\times\exp\left(\sum_v\frac{-\left(-\sum_k\widetilde{\beta}'_{k,v}\log\widetilde{\beta}'_{k,v}\right)^2+\left(-\sum_k\beta'_{k,v}\log\beta'_{k,v}\right)^2}{2\sigma_{l_v}^2}\right)\times\\
&\left.\exp\left(\frac{-(-\sum_k\tau_{kw_{m,n}}\log\ \tau_{kw_{m,n}})^2+(-\sum_k\widetilde{t}_{kw_{m,n}}\log\ \widetilde{t}_{kw_{m,n}})^2}{2\sigma_{l_v}^2}\right)\right)
\end{aligned}\tag{4.14}$$

It can be seen that this expression can be computed in $O(1)$ time by caching $\tau_{kw_{m,n}}$ and $\beta'_{k,v}$ values along with their log-values. Therefore the computational complexity for sampling is the same as collapsed Gibbs sampling — $O(INK)$ (as before, I is the number of iterations, N the number of words in the corpus and K the number of topics). The constant time added while checking for acceptance is however large due to the expensive exponentiation and log functions involved.

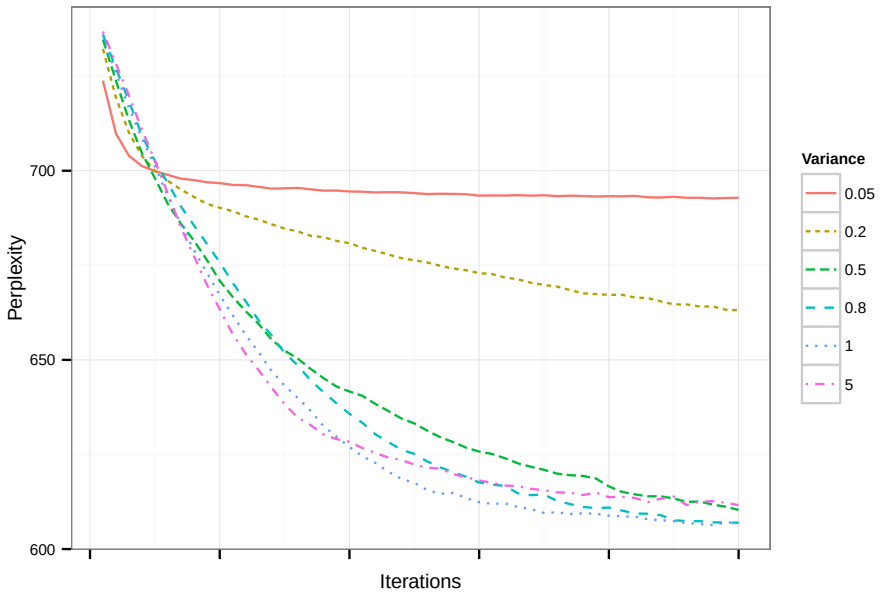


Figure 4.3: Metropolis Hastings sampling - Tracking Perplexity (on a 10% held out test set) vs. Iterations

4.3 Empirical Results

Firstly, we empirically verify that the LDA model with the ECD prior converges to a nearly steady perplexity value. A corpus of 1176 Amazon product reviews with a vocabulary size of 2012 is fit with LDA using 10 topics. Figure 4.3 show the perplexity for a held out subset of the corpus for different values of the $\sigma_{l_v}^2$ hyperparameter. The plots indicate that the model indeed converges to a steady perplexity value. We see that the optimal value for the hyperparameter at which the perplexity is lowest is 1.0.

Next, we track the average word entropy as the iterations progress in Figure 4.4. As expected, the plots show that a higher variance value permits higher entropy in the latent role distributions for words. We also note that the average word entropy levels off relatively early in the inference process as compared to the perplexity.

Using the ECD prior with the Metropolis Hastings sample scheme described is a computationally expensive operation. Equation 4.14 will need to be computed for every word in every iteration. As explained earlier, the acceptance ratio term in the MH algorithm determines if the newly proposed

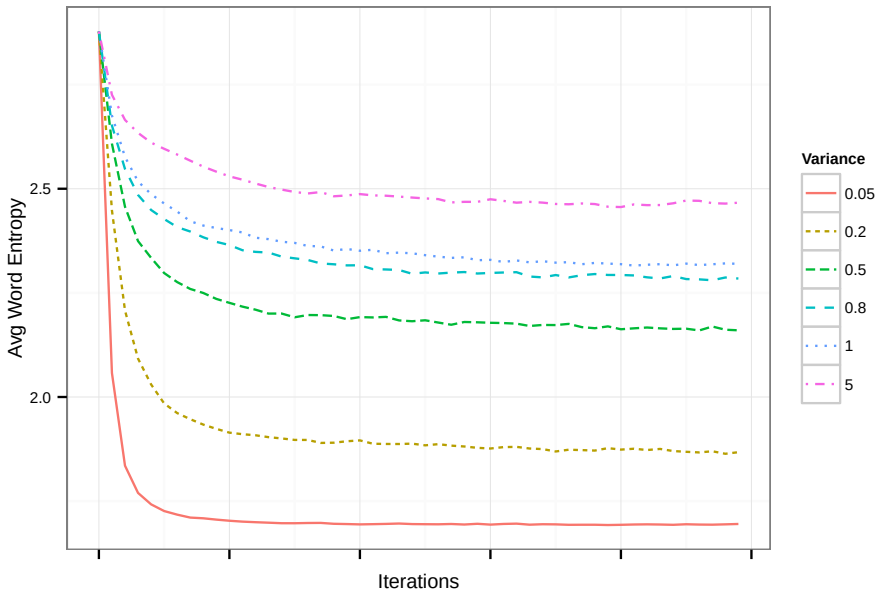


Figure 4.4: Metropolis Hastings sampling - Tracking Entropy (on a 10% held out test set) vs. Iterations

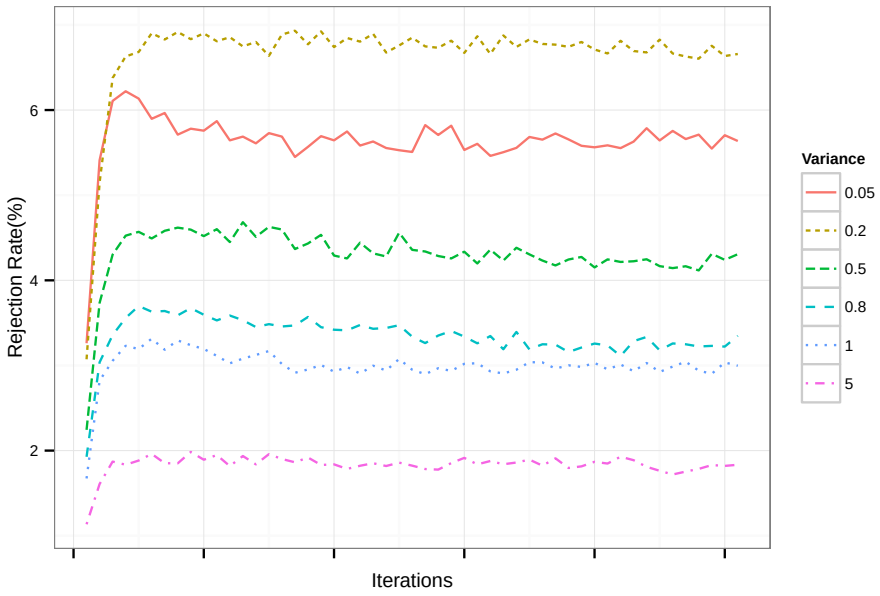


Figure 4.5: Metropolis Hastings sampling - Tracking Rejection Rate (in the training set) vs. Iterations

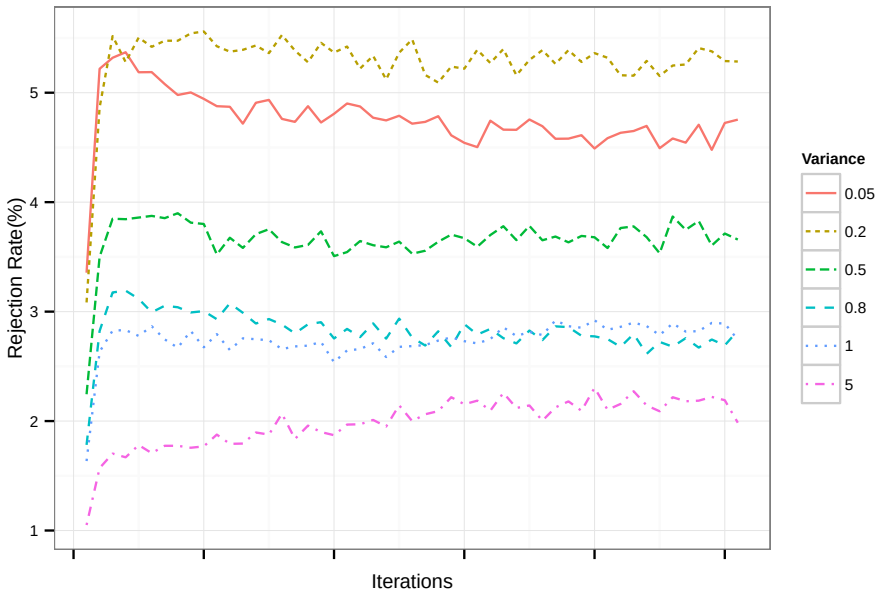


Figure 4.6: Approximated Metropolis Hastings sampling - Tracking Rejection Rate (on train set) vs. Iterations

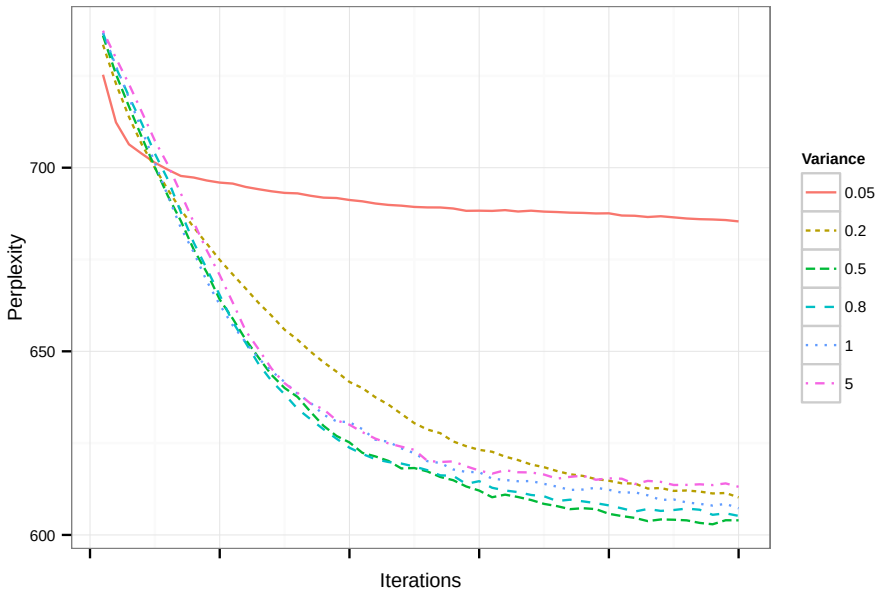


Figure 4.7: Approximated Metropolis Hastings sampling - Tracking Perplexity (on held out test set) vs. Iterations

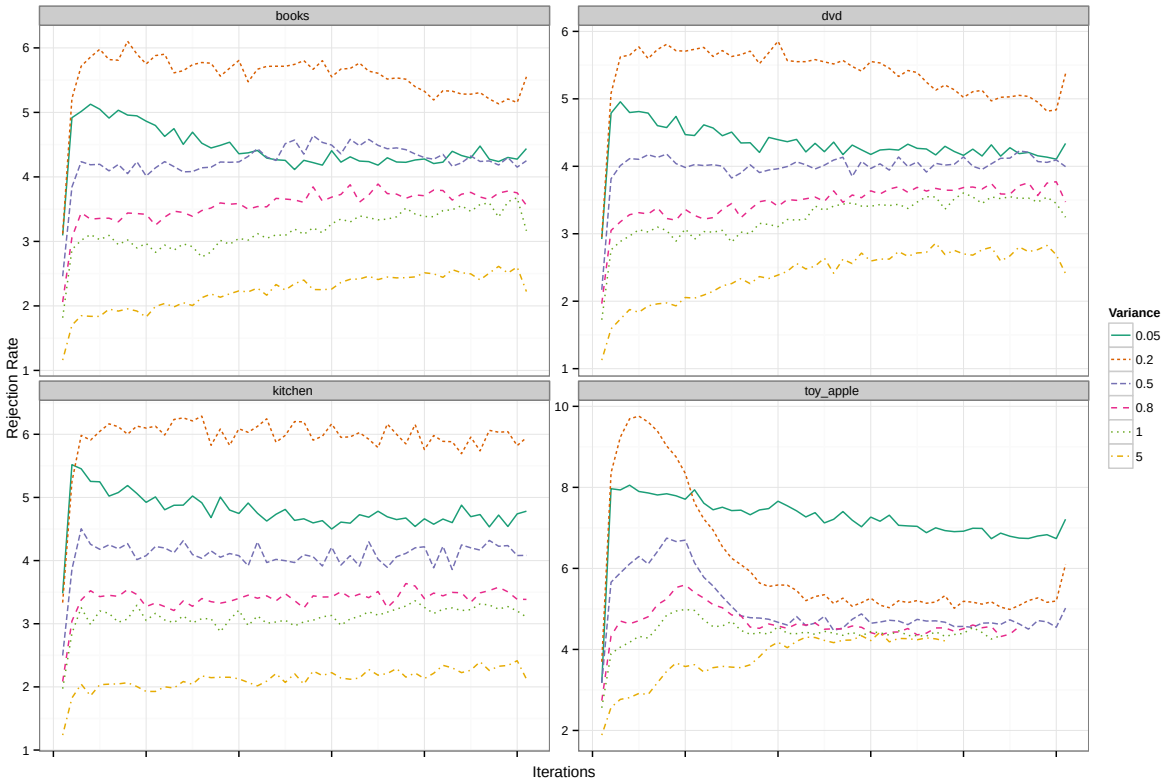


Figure 4.8: Tracking Rejection rate (on heldout set) in other datasets.

sample is accepted or not. If the rate of acceptance is ~ 1, we could potentially accept all new proposals without the need to compute the expensive acceptance ratio. In Figure 4.5, we track the rejection rate of the samples proposed during each iteration of the MH algorithm, The experiment is repeated for different hyperparameter values. The plots indicate that after convergence, an average of 1.9% to 6.4% of samples are rejected, depending on the hyperparameter value used. It can be noted here, that if we use the Metropolish Hastings procedure without rejecting any samples, using the proposal distribution described earlier, the procedure reduces to collapsed Gibbs sampling. Since the rejection rate observed in the above experiment is a relatively low fraction, we propose that the collapsed Gibbs sampler proposed for the regularized LDA model i.e., the proposal distribution for the MH algorithm proposed in this chapter, be used as a proxy for the expensive Metropolis Hastings sampler. Essentially, this means that we accept newly proposed samples unconditionally without actually computing the acceptance ratio.

A possible danger of this approach is that the small fraction of unacceptable proposals could

lead to a drift into undesirable areas in state space. To check for this, we use the collapsed Gibbs sampler but also compute the acceptance ratio when a new topic identifier is chosen (The computed acceptance ratio is however not used for any purpose other than for logging). Figure 4.6 shows that the rejection rate is similar to the values in Figure 4.5 which indicates that the approximation introduced by accepting all samples does not empirically lead the model astray. Figures 4.7 also shows the perplexity trace with the collapsed Gibbs sampler, which is similar to Figure 4.3, which further reinforces the claim that universal acceptance of new samples does not hurt empirical performance while providing computational gains.

In general, we recommend that the decision to use the approximation of always accepting the proposed new sample, i.e., resorting to Gibbs sampling, be done by looking at the rejection rate in the Metropolis Hastings process, for the particular dataset in question. Figure 4.8 shows the trace of rejection rates over iterations on other reviews datasets and an entity dataset (described in Section 5.2). It can be seen that the rejection rates in all these datasets are similar to the rates in figure 4.5, which indicates that avoiding the expensive Metropolis-Hastings acceptance probability calculation is potentially possible in other datasets also.

4.4 Related Work

Wang and Blei [2013] present generic methods to use mean field variational inference in situations where models have non-conjugate priors. This approach is a feasible alternative to the MCMC method presented in this chapter. Jain and Neal [2000] proposed a method to overcome problems in using collapsed Gibbs sampling for mixture models which tend to get trapped in local modes when mixture components are similar. The authors propose a Metropolis Hastings alternative that uses a complex proposal distribution to overcome the problems. Steck and Jaakkola [2002] describe the characteristics of using a product of independent Dirichlet priors in a Bayesian regularization setting in applications where the aim is to retrieve structure rather than fit parameters.

4.5 Conclusion

In this chapter, we presented a Bayesian framework which achieves the same effect of controlling the entropy of words' latent role distributions as the regularizer presented in the previous chapter did. We demonstrate that we can construct a prior distribution for topics and document distributions that mimics the behavior of the regularizer. Since the Bayesian prior prevents us from using collapsed Gibbs sampling, we proposed a Metropolis Hastings inference scheme for a topic model using the newly proposed ECD prior. We also see that the rejection rate during Metropolis Hastings for the datasets used in our experiments was low which allows us to approximate inference by always accepting the proposed samples, thereby providing considerable savings in computational costs.

Chapter 5

Limited Supervision in Topic Models

5.1 Introduction

Topic modeling is typically used in a fully unsupervised setting; as such, it is unequipped to utilize limited supervisory information, e.g., feature labels and document cluster membership. Here, we introduce methods to incorporate progressively stronger forms of weak supervision to influence the formation of topics that respect *apriori* information that we might have about the latent structure.

First, we use the regularization approach presented earlier to to bias mixed-membership models to better exploit known *topic-indicative* features, i.e., features that are strongly indicative of latent topic. To achieve the bias, We need the inference procedure to be able to restrict the freedom accorded to such features to span multiple latent roles. For instance, while using Link-LDA to model academic publications in biology about the yeast organism, mentions of protein names are strongly indicative of topic, i.e., a single protein is much less likely to occur under different topics than other natural language tokens. Unsupervised topic models do not necessarily optimally utilize this kind of *apriori* information. The bias term we introduce serves to control the latent role distribution of the features, i.e., the degree of *polysemy*, and its strength can be adjusted to control the degree of polysemy permitted. The flexibility of the biased models is examined by using it to cluster entities found in HTML pages [Dalvi et al., 2012]. While the approach can be used for a variety of tasks, we focus on the HTML table derived entity clustering task since it requires the use of several kinds of features (obtained from semi-structured data from the tables) and permits us to

demonstrate ways in which intuition and limited supervision about different kinds of features can be incorporated. In this task, potentially useful features of a document (representing features of an entity) include features like the headers of columns (e.g. the entity *apple* might be found under the headers *company* or *fruit*) and web-domains (e.g. food.com, finance.com, etc.) in which it was found. The biasing technique presented could be used to capture our domain knowledge that features of a certain type are more topic-indicative than other features. When the bias term is set high, the features to which it is applied are deemed to be more strongly indicative of topic and are strongly discouraged from assuming multiple latent roles in the mixed membership model.

Next, stronger forms of supervision, in the form of feature and document labels are injected into the model, to obtain models that range from fully unsupervised topic models to semi-supervised models. This form of light supervision can be in the form of known latent roles for certain subsets of topic-indicative features, or known latent roles for single-membership documents (i.e. non-polysemous entities). The supervision is incorporated into the model by modifying the collapsed Gibbs sampling procedure used for approximate inference.

5.2 Entity Clustering

Latent-variable mixed-membership models based on LDA are used for a variety of tasks in NLP. Here, we use a regularized version of Link-LDA [Erosheva et al., 2004] for the task of clustering entities that are extracted from tables in HTML documents crawled from the web [Balasubramanyan et al., 2013]. Dalvi et al. [2012] describe the task in detail.

In this task, the dataset consists of tables of entities extracted from HTML pages. The tables for instance could contain lists of companies, music composers, soccer teams etc. The goal of the task is to cluster entities of the same semantic class together. Therefore, if the dataset includes a table of fruits with apples, grapes and oranges, and another table with oranges, peaches and bananas, the goal of the task is to recover a cluster of fruits which includes apples, grapes, oranges, peaches and bananas.

Surface terms in such HTML tables frequently have multiple senses. For example, consider the term *apple*, which could be found in tables of companies and fruits among others. Therefore,

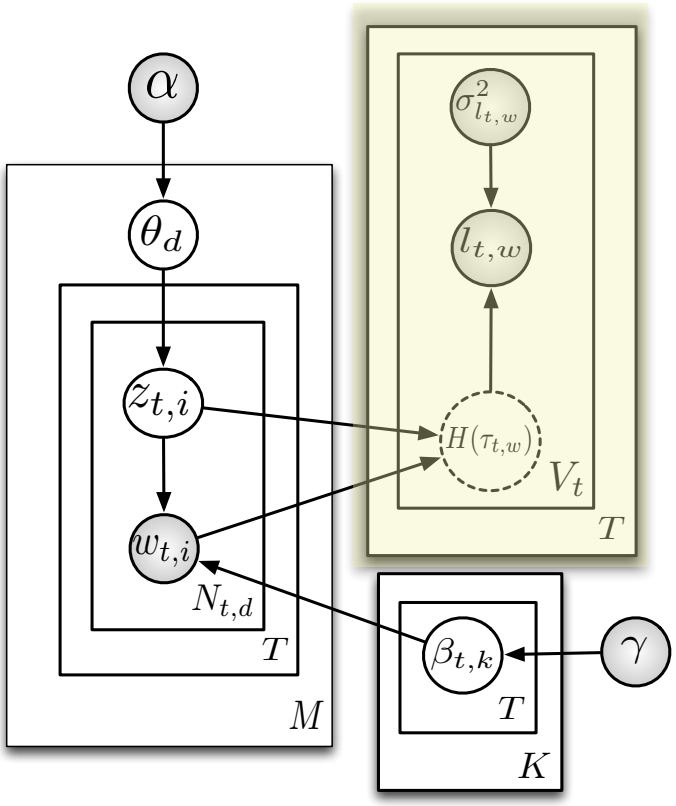


Figure 5.1: Biased Link-LDA model to Exploit Topic Indicative Features.

we need a model that is capable of distinguishing the sense of the term to prevent companies and fruits from being collapsed into one cluster based on the term *apple* frequently co-occurring with both companies and fruits. Mixed-membership models can account for the multiple-sense problem by assigning partial membership in both clusters to the entity. Typically, entity clustering has been based on distributional similarity based approaches or by using text patterns [Hearst, 1992]. In this task however, since we are dealing with entities in HTML tables as opposed to entity mentions in free text, we use a different set of features to assist in the clustering, namely: a) co-occurring entities, b) co-occurring entity pairs, c) the table id-column id pairs under which the entity was observed, d) web domains in which the entity was observed, and e) the hyponyms that are associated with an entity, which are extracted using Hearst patterns from the Clueweb corpus. The task can therefore be seen as distributional clustering with a different set of contextual features than the free text features usually used. For every unique entry found in the collection of tables in

a dataset, we construct a *"document"* in the LDA sense with the above five kinds of *"words"*. The document is represented by a set of bags of words, one for each kind of feature used. A document for the entity *apple*, for example might consists of the following bags - a. co-occurring entities {*orange*, *apple*, *microsoft*, ... }, b. entity pairs {*orange:apple*, *google:apple* ... } c. column ids {tab:326::colid::1 ...} d. domains {business.com, produce.com ...} e. hyponyms {*stocks*, *juice*, *tech companies* ...}. These different classes of features are modeled using the Link-LDA model [Erosheva et al., 2004]. Figure 5.1 shows the plate diagram of the graphical model. The variables that are under the yellow shaded rectangle provide the bias that is introduced by regularization.

Evaluation

The predicted clusters are evaluated using *Normalized Mutual Information (NMI)* [MacKay, 2002]. This score measures the amount of information about correct cluster labels that is encoded by the predicted topic/cluster distributions. NMI can be used in mixed-membership scenarios since it does not require the true cluster distribution and predicted topic distribution to have membership in one single cluster. Additionally, the number of true clusters and topics do not have to be the same and therefore no mapping from latent topics to known cluster labels is required. To compute NMI between the true cluster label distribution and predicted distributions for the test entity set E_{test}, first we compute Ω, the predicted distribution of topics given by

$$\Omega_k = \frac{\sum_{e \in E_{\text{test}}} \theta_{e,k}}{|E_{\text{test}}|}, \quad k \in \{1, \ldots, K\} \tag{5.1}$$

Let $\mathcal{C}$ be the distribution over true cluster labels, i.e,

$$\mathcal{C}_k = \frac{\sum_{e \in E_{\text{test}}} \mathbb{I}(\text{true-label}(e) == k)}{|E_{\text{test}}|}, \quad k \in \{1, \ldots, K\} \tag{5.2}$$

The NMI score for the predicted distribution of the test set E_{test} is given by:

$$\frac{I(\Omega; \mathcal{C})}{\left(H(\Omega) + H(\mathcal{C})\right)/2} \tag{5.3}$$

where I indicates mutual information.

It should be noted that NMI ranges from 0 to 1. Higher NMI scores indicate better performance since it means that the predicted cluster distributions contain more information about the true cluster labels.

Dataset	entities	Size of vocabulary				
		co-occurring entities	entity pairs	column ids	domains	hyponyms
Asia NELL	33455	18309	141352	9477	3207	31833
Clueweb Sports	29113	28891	354614	59117	8088	28618
CSEAL Useful	34565	24340	217328	7337	2118	28381
Delicious Music	18074	9748	106401	7564	1633	24934
Delicious Sports	6786	3183	24147	2050	509	16380
Toy Apple	2411	423	4737	109	53	2826

Table 5.1: Dataset Statistics

While the model returns mixed-membership assignments for entities, the human labeling scheme used in experiments below provides only one true cluster assignment for an entity. We however present a qualitative analysis of the advantages of mixed-membership modeling in Section 6.5.

Entity clustering experiments were performed using the WebSets datasets [1] [Dalvi et al., 2012], namely the Asia NELL, Clueweb Sports, CSEAL Useful, Delicious Music, Delicious Sports and Toy Apple datasets. The Asia NELL dataset was collected using the ASIA system [Wang and Cohen, 2009] using hypernyms of NELL [Carlson et al., 2010a] entities as queries. The Clueweb Sports dataset consists of tables extracted from Sports related pages in the Clueweb dataset. The Delicious music and sports datasets consist of tables from subsets of the DAI-Labor [Wetzker et al., 2008] Delicious corpus that were tagged as music and sports respectively. The Toy Apple dataset is a small toy dataset constructed using the SEAL [Carlson et al., 2010b] system to create set-expansion lists using the query "Apple", which is a typical example of a multi-sense entity (as a fruit and as a company). It is used primarily to illustrate the effects of clustering mixed membership entities. The hyponyms features for all datasets were extracted using Hearst patterns on the Clueweb dataset. Statistics about the datasets are shown in Table 5.1.

[1] http://rtw.ml.cmu.edu/wk/WebSets/wsdm_2012_online/index.html

5.3 Exploiting Topic Indicative Features using Regularization based Biased Models

One of the attractive attributes of topic models is that they require no supervision in terms of data annotation. However, in many situations, we might have some apriori information in the form of intuition about which features are topic-indicative. We use the entropic regularization approach presented in previous chapters to bias topic models to utilize weak knowledge about features. Specifically, we aim to make the model exploit *topic indicative features*, which are known to be strongly indicative of topic. For instance in the toy apple example, the co-occurring entities features of the ambiguous entity *apple* are topic indicative. Co-occurring entities such as *Google* and *Microsoft* are indicative of the company topic where as co-occurring entities like *grape* and *banana* indicate the fruit topic. The bias is introduced into the model via a regularization term that constrains the freedom of specific features to take on multiple latent roles. This results in models where every unique word in the vocabulary is strongly indicative of a latent topic. When the bias is turned up in strength , i.e., as $\sigma_{l_w}^2$ tends to 0, the inference procedure effectively partitions the vocabulary into subsets corresponding to the different latent topics.

5.4 Injecting Labeled Features and Documents

In this section, we study how stronger prior knowledge in the form of labeled features and labeled documents can be incorporated into mixed-membership models. *Topic tables* (e.g. Table 3.2) are a commonly used method to display latent topics that are uncovered using models such as LDA. These tables depict topics using the top words of multinomials recovered after inference. Here, we use labeled features to indicate the topic a feature belongs to. as a way to influence the formation of the topic tables. This is done by giving the model hints about the latent topic tables that we expect to see for the labeled features. Document labels, similarly bring the model closer to semi-supervised learning where a subset of the training data has known labels by providing *apriori* information about the latent topic assignment during inference.

As a concrete example, let us return to the task of clustering entities drawn from web tables.

We might have domain knowledge that certain entities do not have multiple senses and should be assigned to a single pre-known latent cluster. An example is *Google* which in the context of our task is known to always be generated by the company topic. In general, we have pre-known latent cluster assignments for a small set of features which are strongly topic-indicative obtained from an expert (usually via human labeling).

While the motivation in using a LDA-derived approach for the entity clustering task lies in its ability to model mixed-membership-ness, in the task of clustering entities, there are many entities that belong to a single cluster. In such a context, it would be useful to allow the inference procedure to use known cluster assignments for a small number of documents to influence the latent cluster formation. For instance, in the entity clustering task using the Toy Apple dataset, we might wish to use domain knowledge to say that "persimmon" belongs exclusively to the "fruit" cluster.

Let $\mathcal{L}$ be a set of pairs $\langle w, k_w \rangle$ where w is a feature i.e. $w \in V_t, t \in 1 \dots T$ and $k_w \in 1, \dots, K$. Each such pair indicates that the latent topic that generates an instance of w in the corpus is almost certainly k_w. Note that we do not have information about the nature of topic k_w before inference. We use the topic ids in $\mathcal{L}$ to separate and funnel features of different known clusters to different topics.

Similary, to formalize the concept of labels for documents, for each labeled documents d in the labeled set D_l, let d be a document that is known to belong to cluster c_d.

Now, we present the generative process for LDA where a subset of features and documents have labels as described in $\mathcal{L}$ and D_l:

1. Generate topics: sample $\beta_{t,k} \sim \text{Dir}(\gamma_t)$ for $t \in 1, \dots, T$, $k \in 1, \dots, K$
2. Generate documents: For each document $d \in D$
   (a) Sample $\theta_d \sim \text{Dir}(\alpha)$
   (b) For each type of entity $t \in \{1, \dots, T\}$
      i. For each instance of an entity $w_{t,i}, i \in \{1, \dots, N_{d,t}\}$
         A. if document d has a known label c_d, set $z_{t,i} = c_d$ with probability γ_d, else
         B. if $w_{t,i}$ has known label k_w, set $z_{t,i} = k_w$ with probability γ_f, else
         C. Sample a topic $z_{t,i} \sim \text{Multinomial}(\theta_d)$

D. Sample $w_{t,i}$ from $\beta_{t,z_{t,i}}$

Due to the conjugacy between the Dirichlet prior for topic multinomial distributions, using labeled features is mathematically equivalent to using different asymmetric Dirichlet priors for each topic instead of the same symmmetric Dirichlet prior, parameterized by γ used previously. For instance if $w \in V_t$ has a label k_w, then the prior for topic k_w is an asymmetric Dirichlet with parameters γ for all words other than w and a larger value γ^* for the word w. For all the other topics, the asymmetric Dirichlet has a lower value $\gamma^{'}$ for w to enforce our prior belief that w is more likely to be generated by topic k_w than any other topic. γ_f is therefore proportional to $\frac{n_{kw}+\gamma^*}{n_k+(V-1)\gamma^{'}+\gamma^*}$. We set the γ_f parameter directly in our approach for better interpretability.

Similar to the case with feature labels, the use of labeled documents is mathematically equivalent to a scenario where labeled documents' topic distributions θ, are drawn from asymmetric Dirichlet priors with higher parameter values for their topic labels instead of the symmetric Dirichlet priors that are usually used. In the implied asymmetric Dirichlet prior used for document d, the hyperparameter for topic c_d, i.e., γ_{d,c_d} has a significantly higher value than the the hyperparameters for the remaining topics.

The collapsed Gibbs sampling process used for inference for LDA needs to be modified to account for the updated generative process. During collapsed Gibbs sampling, when the topic indicator for a word is inferred, the procedure is modified to include a check to see if the word in question is present in $\mathcal{L}$. If yes, then instead of sampling a topic indicator for the word, the latent topic indicator is set to k_w with a probability of γ_f, where γ_f is a constant close to 1. Similarly, to account for document labels, for all words in the document, the cluster c_d is assigned with probability γ_d (≈ 1.0), and the usual collapsed Gibbs sampling procedure is used to determine the latent topic assignment with probability $1 - \gamma_d$.

5.5 Experimental Results

First, we study the effect of biasing the model to better exploit topic-indicative features. As described in Section 5.2, we use regularized Link-LDA to model entities extracted from tables in HTML pages. First we evaluate the model by studying the perplexity of co-ocurring entities

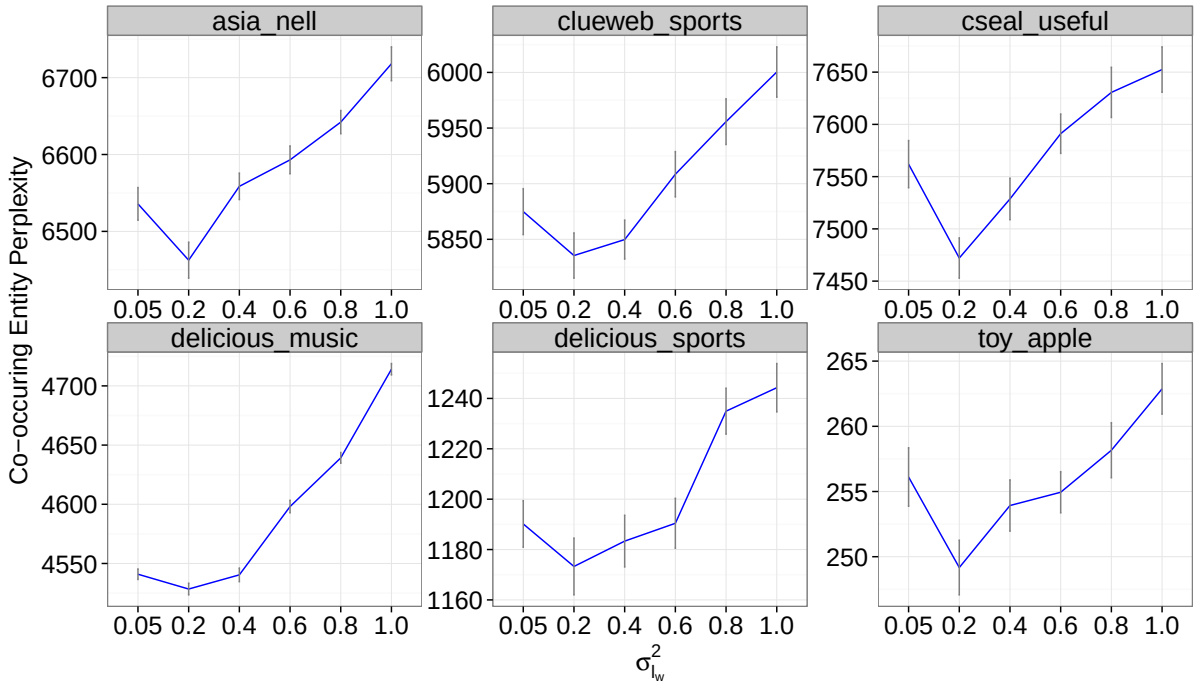


Figure 5.2: Studying perplexity with feature regularization (on 10% heldout tuning dataset)

which is one of the types of features used to represent a entity. Figure 5.2 shows the co-occurring entities perplexity of the biased Link-LDA model for the different datasets, for different values of the variance parameter $\sigma^2_{l_w}$ in the bias term. The reported values are averaged over 10 trials. For each trial, the collapsed Gibbs sampler was run for 100 iterations. The number of topics is set to 50 based on cross-validation. It can be seen that the best perplexity is seen across all datasets when the variance is set to 0.2. We use this variance when using feature regularization, i.e. biasing, for the rest of the experiments. When biasing is used, it is applied to the column id and entity-pair features: a column in a table is unlikely to contain entities from multiple clusters and is therefore strongly indicative of the topic; similarly, while an entity can belong to multiple topics, an entity-pair such as "apple:peach" is strongly indicative of a single topic.

Table 5.2 shows the difference in performance between the biased and baseline unbiased models as measured by NMI between predicted cluster distributions and known true cluster labels of labeled documents. For all the datasets, the biased models show a noticeable improvement over the unbiased variant. Figure 5.3 shows the sensitivity of NMI to the hyperparameter value. We see

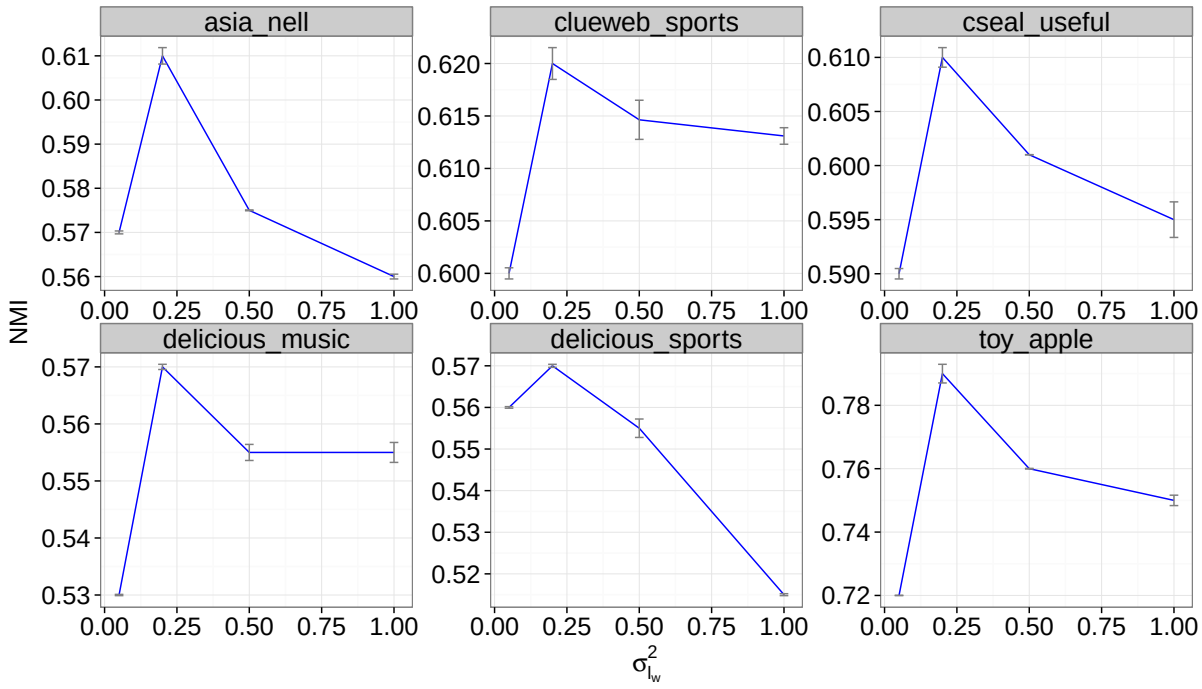


Figure 5.3: Studying NMI with feature regularization (on heldout labeled test set)

Dataset	Regularization		Change
	No	Yes	
Asia NELL	0.586	0.637*	+8.70%
Clueweb Sports	0.567	0.624*	+10.05%
CSEAL Useful	0.533	0.588	+10.31%
Delicious Music	0.548	0.621*	+13.32%
Delicious Sports	0.609	0.615*	+0.98%
Toy Apple	0.771	0.781*	+1.29%

(* - statistically significant at the 0.05 level)

Table 5.2: Feature regularization: Effect on NMI (computed using 10-fold cross-validation)

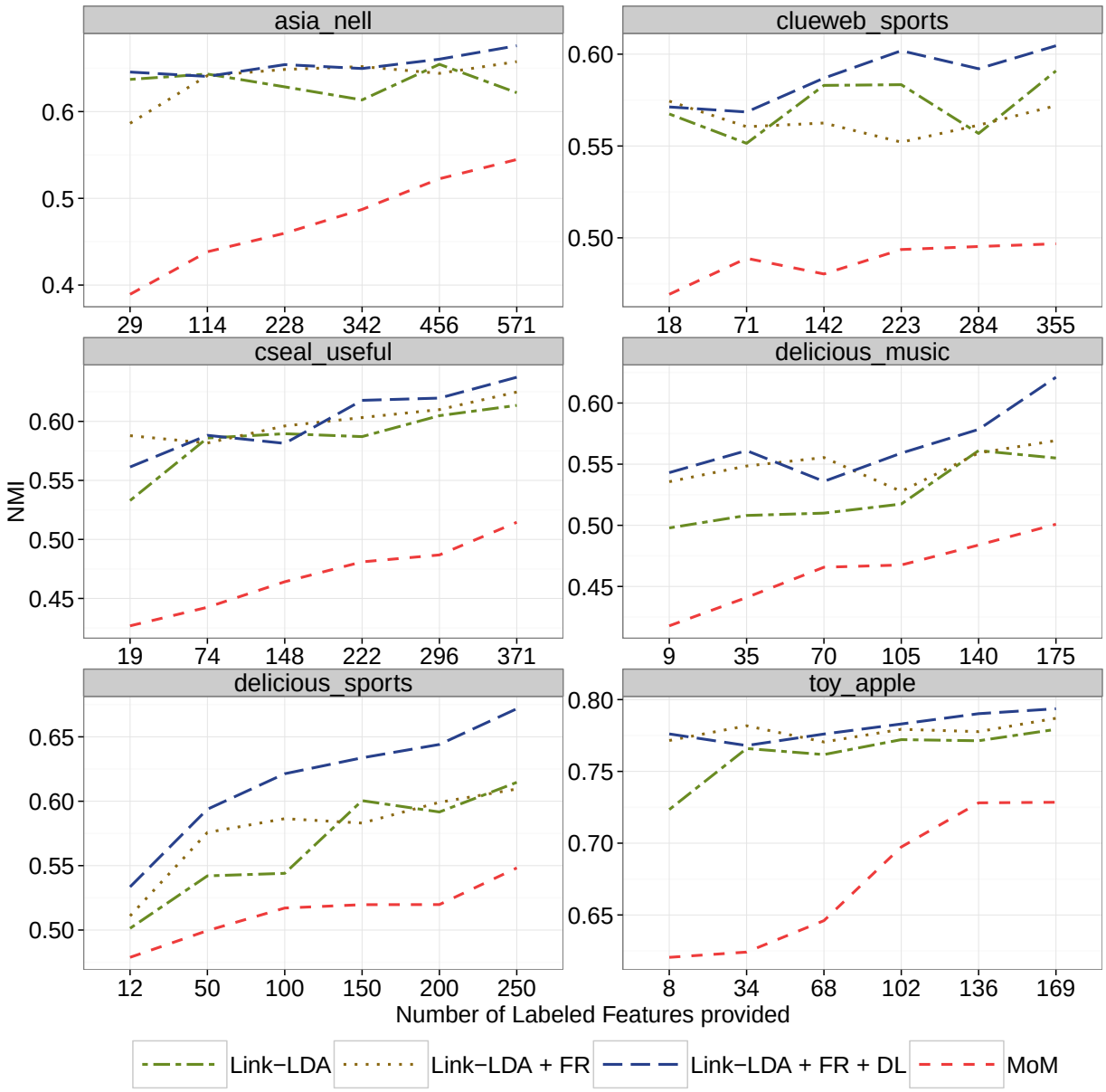


Figure 5.4: Effect of injecting Feature Labels (NMI computed using 10-fold cross-validation)

from the plot that the NMI values are correlated with the perplexity values seen in figure 5.2.

Next, we study the effects of feature and document labeling in Figures 5.4 and 5.5. Feature and document labels are provided to the model for a subset of co-occurring entity features and entities. Labels for entities were obtained using Amazon's Mechanical Turk and were used to label entity documents and also co-occurring entity features. Although entities in general may have

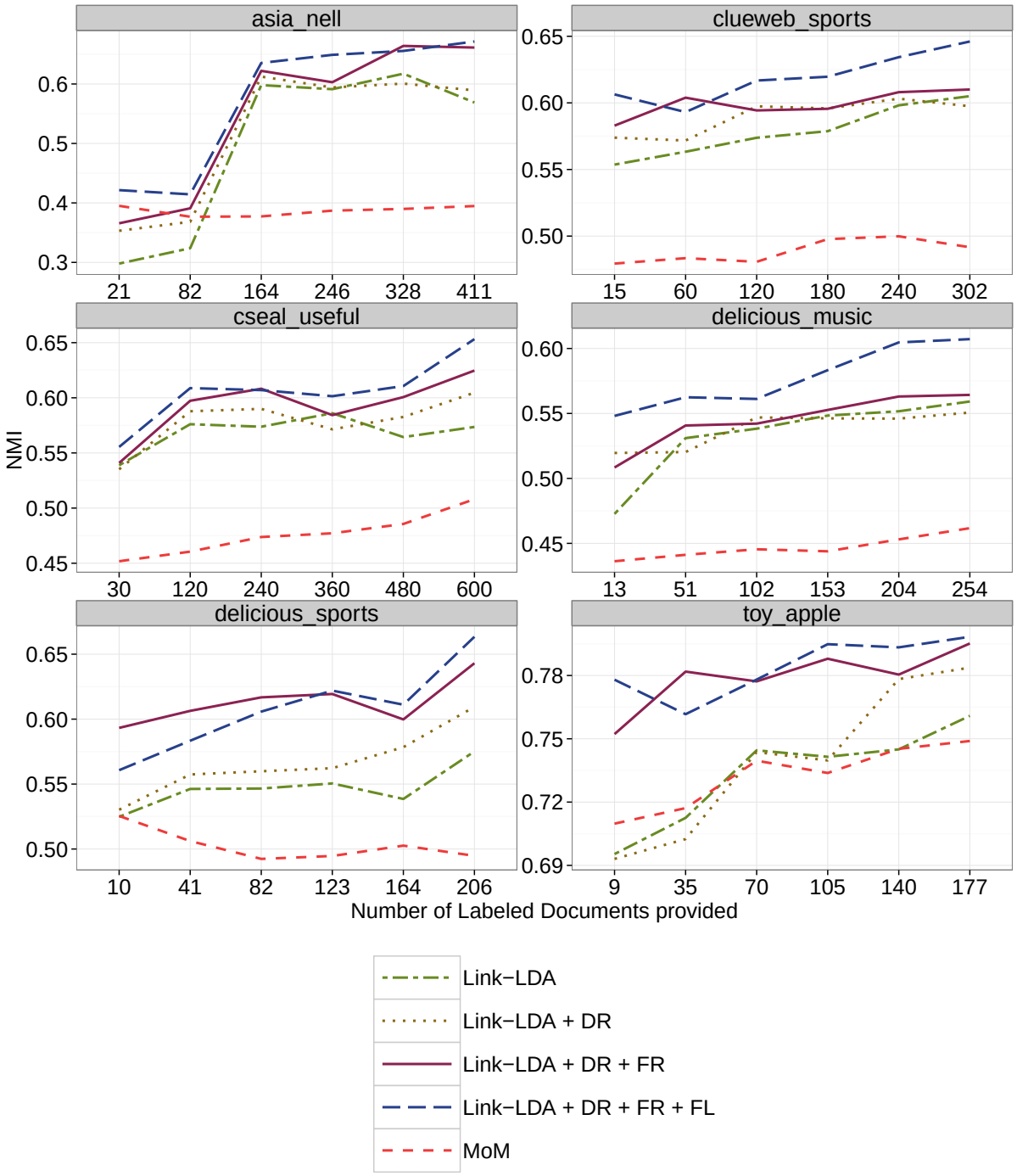


Figure 5.5: Effect of adding Document Labels (NMI computed using 10-fold cross-validation)

multiple senses, we only obtained labels for entities that have a single dominant sense. Table 5.3 shows the number of labeled features and documents for each dataset. In these figures, models

Dataset	Co-occurring entities vocabulary size	#Labeled features	#Labeled documents
Asia NELL	18309	571	411
Clueweb Sports	28891	355	302
CSEAL Useful	24240	371	600
Delicious-Music	9748	175	254
Delicious-Sports	3183	249	206
Toy Apple	423	169	177

Table 5.3: Feature and Document Label statistics

are trained with increasing amount of supervision in the form of feature and document labels and the NMI between the true cluster labels of labeled documents and their inferred topic distributions for different model variants are plotted. It can be seen that as expected, increasing the amount of labeled data provided to the model results in higher NMI values for all model variants.

The entropy of θ can be subject to the same kind of regularization as the word topic distribution used in feature regularization, enabling us to restrict the degree to which entities are allowed to exhibit mixed-membership. In figure 5.5, it can be seen that adding such document regularization (+ DR), shows better performance than the regular Link-LDA model. It should be noted that we can add both feature and document regularization to the model simultatenously. When adding feature biasing (+ FR) and all available feature labels (+ FL), along with different degrees of document labels, we see a progressively higher NMI across all datasets especially as the number of labeled documents provided is higher. The red dashed line in the plot representing the performance of the MoM model, shows the performance of a single cluster membership model as we move from a fully unsupervised model to a semi-supervised model on the right. It can also be noted here that the green long dashed line representing Link-LDA is equivalent to using the Labeled-LDA model when document labels are added. Labeled-LDA [Ramage et al., 2009] works by constraining a document's topic distributions to only have mass on pre-specified topics for each topic. By using document labels, we achieve a similar effect, with the additional restriction that only one label is provided for a document.

In figure 5.4, the red dashed line shows the performance of a mixture of multinomials (MoM) model [2] which allows each entity to belong to exactly one cluster. It can be seen that disallowing mixed-membership results in lower performance as compared to even the plain vanilla LDA model. The plot also indicates that the adding feature regularization (Link-LDA+FR) i.e. biased features consistently shows higher NMI values than the unbiased Link-LDA model and that adding all available document labels (Link-LDA+FR+DL) in addition to the different amounts of feature labels to the biased Link-LDA model yields the best NMI.

The above experiments show that introducing labeled documents and features consistently improves performance. While document labels have more impact, the labeling scheme used restricts us to only provide labels for entities with a single sense. We also see that for a fixed number of feature or document labels, adding feature regularization (i.e. biasing) and document regularization consistently improves the NMI scores.

Table 5.4, shows illustrative examples of the advantage of the mixed-membership approach. For the ambiguous entities shown, the top two topics to which they are deemed to belong are shown using the top entries from the entity-pair multinomials. The results are from a biased Link-LDA model with no labeled features or documents. The topic titles in bold were added after inference by looking at the top entries for the topic. The value in parentheses show the degree of membership that the entity has for the topic. It can be seen that the mixed-membership latent variable model approach is able to detect the multiple senses of ambiguous entities. The first entity in the table **franklin** is ambiguous because it has multiple senses — as a common first or last name and as a name of a city in the state of Nebraska in the US, among others. The second example **apple** as discussed earlier could either refer to the fruit or the company. The top two topics returned for this entity denotes exactly these two concepts. The third example **giants** is from the sports domain and could refer to either the New York Giants who play in the National Football League (American Football) or the San Francisco Giants who play in Major League Baseball (MLB). The top two topics indicate these two concepts.

We note here that we cannot quantitatively compare the entity clustering results from these experiments to the results from prior work in HTML table based entity clustering by the WebSets

[2]While EM can be used for inference in the MoM model, we use collapsed Gibbs sampling for these experiments.

Dataset: asia_nell, Entity: **franklin**
Names: (0.24) armstrong:brown, jennifer:jessica, chloe:gucci, brandon:joseph, benjamin:matthew, donald:edward, russell:stanley, benjamin:ethan, greg:gregg, angel:jose
Places: (0.21) montana:nebraska, dakotas:north_carolina, rock_island:san_francisco, atlanta:long_island, delaware:montana, montana:new_york, central_california:san_clemente_island, clearwater:cocoa_beach, sutter:tehama, oklahoma_city:salt_lake_city
Dataset:toy apple, Entity: **apple**
Food: (0.61) peaches:pears, cocoa:coconut_oil, apricots:avocados, sodium_carbonate:sodium_chloride, lactic_acid:lauric_acid, sugar_alcohols:sugars, coconut_oil:coffee, caffeine:calcium_carbonate, xanthan_gum:yeast, sodium_citrate:sodium_hydroxide, pears:pineapple
Companies: (0.16) nec:palmone, blackberry:google, sony:tomtom, asus:palm, philips:samsung, dell:ericsson, sagem:sharp, orange:philips, asus:google, sagem:samsung, asus:bosch
Dataset:delicious sports, Entity: **giants**
NFL teams: (0.26) chiefs:redskins, browns:raiders, cardinals:redskins, rams:saints, cowboys:redskins, cowboys:jaguars, bengals:eagles, bengals:patriots, falcons:patriots, saints:falcons, eagles:panthers
MLB teams: (0.21) arizona_diamondbacks:cincinnati_reds, pittsburgh_pirates:texas_rangers, cleveland_indians:minnesota_twins, milwaukee_brewers:san_diego_padres, boston_red_sox:los_angeles_dodgers, cincinnati_reds:new_york_yankees, minnesota_twins:pittsburgh_pirates, florida_marlins:houston_astros, chicago_cubs:los_angeles_dodgers, baltimore_orioles:montreal_expos, houston_astros:philadelphia_phillies

Table 5.4: Mixed-membership clustering results of ambiguous entities.

system [Dalvi et al., 2012], because the approach in that work clusters triplets of entities extracted from tables rather than individual entities. However, one qualitative difference is that the biasing

technique presented here is a general one and can be applied to any task that mixed-membership models are used for, whereas the WebSets approach specifically addresses the entity clustering task. For rough comparison, however, the NMI value of clustering entity-triples from the Delicious-Sports dataset is reported at 0.64 using the WebSets approach whereas Table 5.2 indicates that the regularized model returns a NMI of 0.615 for the same dataset.

5.6 Related Work

Incorporating document labels into classifiers to obtain semi-supervised models is a well established technique in machine learning [Nigam et al., 2000]. In the context of topic models, Labeled-LDA [Ramage et al., 2009] uses tags attached to documents to limit the membership of the documents to specified topics. The labeled document injection technique discussed in this chapter is closely related to Labeled-LDA. Rubin et al. [2012] present a set of models, one of which - Flat-LDA is similar to Labeled-LDA and the document label concept presented in this paper. Flat-LDA requires all documents to have labels which our framework does not require, but permits documents to have multiple labels. Wang et al. [2007] presented Semi-Latent Dirichlet Allocation, where the latent topic indicators for words are known a-priori. This approach is similar to the feature labeling framework described in this chapter. Their approach also permits a word to assume different known latent roles in different instances. A more detailed exposition of this approach was presented by the authors in a later paper [Wang and Mori, 2009]. Andrzejewski and Zhu [2009] describe a setting for incorporating labels for features where labels are assigned to words on a per-instance basis rather than for all instances as in our approach. An extension of this work where topics are drawn from Dirichlet forests was presented in Andrzejewski et al. [2009]. Mao et al. [2012] present a method to incorporate hierarchical labels of documents in a hierarchical LDA setting. A similar method to model documents organized into taxonomies was presented in Bakalov et al. [2012]. Both these approaches are closely related to the document labeling framework described in this chapter.

Steyvers et al. [2011] present an approach where topics are "pre-constructed" based on concepts obtained from Cambridge Advance Learner's Dictionary (CALD). This approach is similar to the labeled features idea presented in this chapter. A concept topic as defined by this approach can be

seen as a set of labeled words with the same topic indicator. Supervised LDA [Blei and McAuliffe, 2008] is a related model where supervision in the form of categorical or real-valued attributes of documents is provided. These attributes are derived from the topic distributions using regression models, which differs from the approach in this paper where the document labels directly indicate topic membership. Zhu et al. [2012] presented MedLDA that like Supervised LDA, predicts attributes for documents, but uses a max-margin approach for the prediction. Mimno and McCallum [2008] proposed a model where the Dirichlet prior for document topic proportion distribution is replaced with a log-linear prior that permits the distribution to be directly influenced by metadata. This work can be interpreted as a method to use metadata to tailor the latent structure formation. Settles [2011] presented the DUALIST system which used labeled features for multinomial Naive Bayes classifiers. This approach is similar to the one employed in our experiments especially when feature labeling is added to the mixture of multinomials model. A similar approach was also used by Attenberg et al. [Attenberg et al., 2010] in the context of active learning. While labeled features have been used in supervised classifiers, they have not been used in latent variable models as far as we know.

Entity clustering from semi-structured data has been addressed previously [Dalvi et al., 2012, Pasca and Van Durme, 2008, Talukdar et al., 2008]. These approaches however do not address the issue of mixed-membership.

Summary

The approach presented in this chapter to introduce limited supervision into mixed-membership models complements the regularization approach presented in the previous chapter. Figure 5.6 shows the landscape of mixed-membership models in three axes *vis a vis* supervised vs. unsupervised models, multi-attribute vs. single attribute models and mixed-membership vs. single-membership models. The document topic proportion regularization approach introduced in chapter 3.3 provides the freedom to traverse the y-axis from a mixture of multinomials model to LDA. The ability to use documents labels for a subset of the corpus as presented in this chapter allows traversal along the x-axis to get models ranging from LDA to Labeled LDA [Ramage et al., 2009]. The approaches

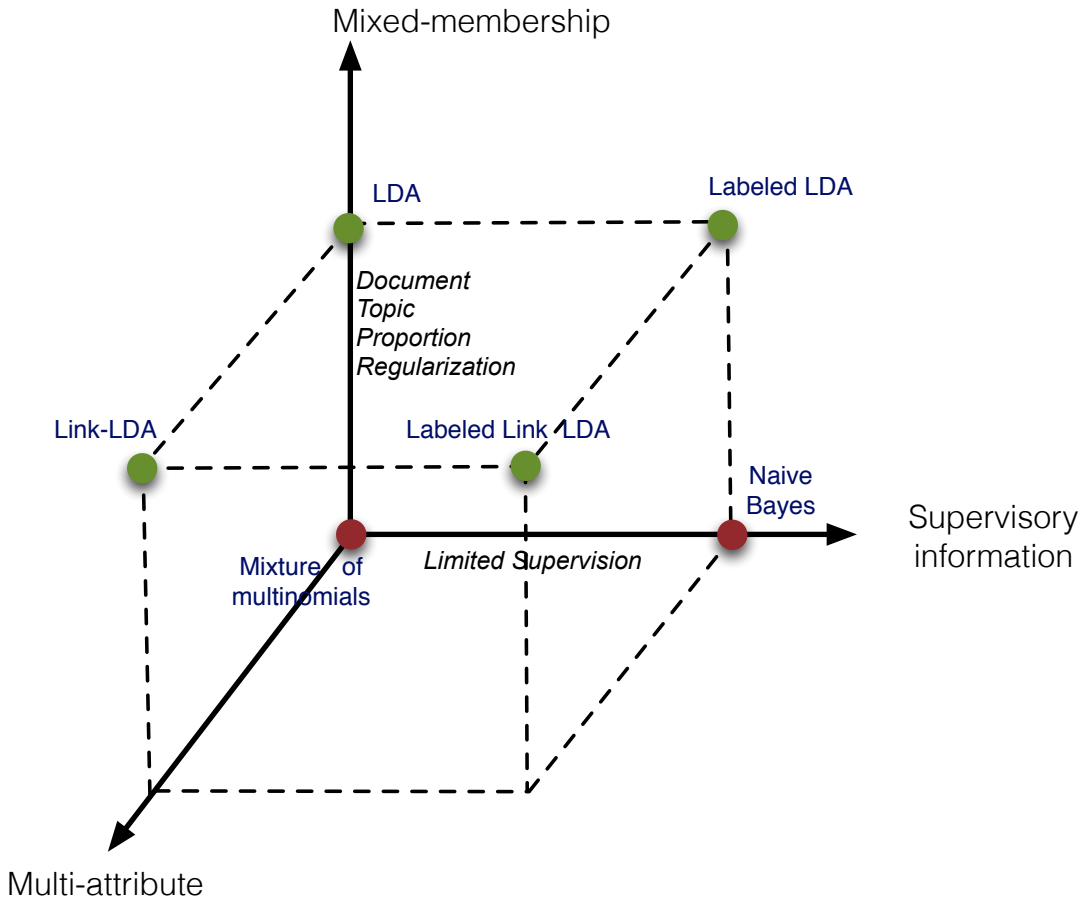


Figure 5.6: Mixed-membership models and their characteristics

presented in this chapter and the previous chapter therefore provide a smooth seamless manner to traverse the range of models seen in figure 5.6.

5.7 Conclusion

We presented a technique to bias latent variable mixed-membership models to exploit topic-indicative features and used the biased model for the task of clustering semi-structured data in the form of entities extracted from HTML tables. Our experiments show that the biased models outperform the baseline models in the cluster recovery task as measured by NMI. We then

presented a method to allow for stronger supervision in the form of feature and document labels to move further along the spectrum toward semi-supervised learning from totally unsupervised learning. Results indicate that the stronger forms of supervision result in better cluster recovery. To summarize, we presented a framework in which mixed-membership models can be successfully used in a semi-supervised fashion to incorporate inexpensive weak prior domain knowledge. The techniques presented in this chapter were published earlier in Balasubramanyan et al. [2013].

Chapter 6

Entropic Regularization in Network Models

6.1 Introduction

Modeling relations between pairs of objects, often by representing them as graphs with nodes corresponding to objects, is a frequently encountered setting in machine learning and statistics. Common examples of such graphs are web graphs, where the relations indicate hyperlinks between webpages; and social networks, with nodes representing people and edges representing a social link between pairs of people. Models of relational data serve as a foundation for tasks like clustering (i.e., grouping nodes by similarities in interaction patterns), de-noising network representations, and visualizing large complex networks.

The task of studying network structure has been a fertile area of research. Here, we mainly focus on stochastic models [Goldenberg et al., 2010, Holland et al., 1983, Snijders and Nowicki, 1997], i.e., generative models that produce random graphs. Stochastic models are a type of network model which posit that nodes play a single latent role and the probability of an edge depends only on the latent roles of the nodes. While this approach is simple and elegant, nodes in complex graphs often exhibit multiple latent roles. For instance, in a social network, a person might assume a personal role while creating a link with a relative or a family member and don a more professional role while doing the same with a colleague. Airoldi et al. [2008] introduced the Mixed Membership

Stochastic Block model (MMSB) which models this phenomenon. This idea of mixed membership shares the same motivation as language models such as Latent Dirichlet Allocation (LDA) [Blei et al., 2003], where a word is free to take on different latent roles when it appears multiple times in the corpus. An alternate network model (henceforth the PSK model) presented by Parkkinen et al. [2009] models sparse networks more efficiently.

Here, we use the regularization approach used for language models (described in Chapter 3) to regularize mixed membership network models, demonstrated using the PSK model, which has been shown to outperform MMSB [Balasubramanyan and Cohen, 2011, Parkkinen et al., 2009] (evaluated using cluster recovery) on sparse networks, to obtain *slightly mixed membership* stochastic blockmodels. As with LDA, we extend the model to include a noisy copy of an aggregate function over latent variables (e.g., the entropy of latent role distributions).

We consider two applications of pseudo-observations based regularization. First, we constrain the latent role membership distributions of nodes by penalizing high entropy. By varying the noise model associated with the copy process for the aggregate variables, one can obtain any desired degree of mixed membership, ranging from a fully mixed membership block model (such as the PSK or the MMSB models) to a classical non-mixed network model.

The second application of pseudo-observed variables is motivated by spectral clustering [Luxburg, 2007, Shi and Malik, 2000], a widely used class of techniques for clustering nodes in a graph, and in particular by the Normalized Cut technique. This method strives to produce clusters that are balanced in terms of the cluster *volumes*. (Here *volume* is the sum of degrees of nodes belonging to the cluster). In this spirit, we propose a second regularization term for mixed membership stochastic models that imposes a preference on balanced volumes.

Results of experiments show that adding either of these penalty terms, or their combination, is beneficial, as measured by the average accuracy in recovering cluster labels in networks with known cluster labels for nodes.

6.2 Sparse Network Model

Airoldi et al. [2008] introduced mixed membership stochastic blockmodels (MMSB) that permit

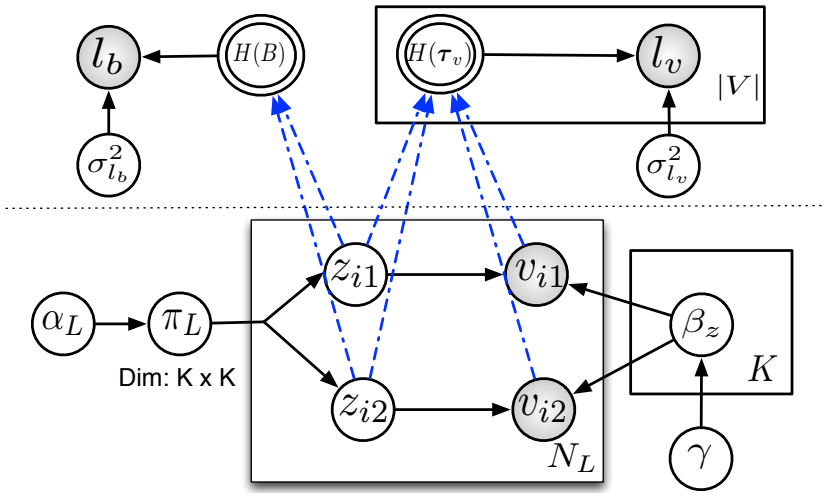


Figure 6.1: The sparse network model with role entropy and volume entropy regularization.

nodes in a network to perform different latent roles in different interactions. The generative process in the MMSB model for a network with V nodes is as follows:

- For each node $v \in V$, draw a K dimensional membership vector $\pi_v \sim$ Dirichlet(α).
- For every pair v_i, v_j, where $v_i, v_j \in V$:
  - Draw a membership indicator for the initiator $z_{i \to j} \sim$ Multinomial(π_{v_i})
  - Draw a membership indicator for the receiver $z_{j \to i} \sim$ Multinomial(π_{v_j})
  - Sample the value of their interaction $Y(v_i, v_j) \sim$ Bernoulli$(z'_{i \to j} B z_{i \leftarrow j})$.

B is the matrix of Bernoulli rates where $B_{i,j}, i, j \in 1, \ldots, K$, indicates the likelihood of a link between nodes assuming latent roles i and j.

The sparse network model (the PSK model) introduced by Parkkinen et al. [2009] is an alternate block model that also allows nodes to take on different latent roles in different interactions like the MMSB model. As in LDA, clusters in this model are treated as multinomial distributions over nodes. Recent literature [Balasubramanyan and Cohen, 2011, Parkkinen et al., 2009] suggests that this model is more suitable when modeling sparse networks.

Figure 6.1 shows the plate diagram for the regularized version of the PSK model that generates a graph representing links between nodes with an underlying block structure. The top part of the figure above the dotted line shows variables related to the regularization. Clusters in this model are

represented as distributions over nodes. Nodes participating in an edge are generated from cluster specific node distributions conditioned on the cluster pairs sampled for the edge. Cluster pairs for edges (links) are drawn from a multinomial defined over pairs of cluster labels. Each node in the set of nodes V in the graph therefore has mixed memberships in clusters. The generative process to obtain links in a graph with K clusters is as follows.

1. Generate cluster distributions:

For each cluster $k \in 1, \ldots, K$, sample $\beta_k \sim$ Dirichlet(γ), the cluster specific multinomial distribution over nodes.

2. Generate edges between nodes:

(a) Sample $\pi_L \sim$ Dirichlet(α_L) where π_L denotes the multinomial distribution over cluster pair labels $\langle k_i, k_j \rangle$, $k_i, k_j \in \{1, \ldots, K\}$.

(b) For every link $v_{i1} \rightarrow v_{i2}$, $i \in \{1 \cdots N_L\}$, where $v_{i1}, v_{i2} \in V$:

(i) Sample a cluster pair $\langle z_{i1}, z_{i2} \rangle \sim$ Multinomial(π_L)

(ii) Sample $v_{i1} \sim$ Multinomial($\beta_{z_{i1}}$)

(iii) Sample $v_{i2} \sim$ Multinomial($\beta_{z_{i2}}$)

In contrast to MMSB, this model only generates realized links that are observed, allowing for better fits for sparse graphs.

Given the hyperparameters α_L and γ, the joint distribution over the links, the cluster pair distribution, the cluster node distributions and cluster assignments for edges is given by

$$\begin{aligned} \mathcal{L} =& p(\pi_L, \boldsymbol{\beta}, \langle \mathbf{z_1}, \mathbf{z_2} \rangle, \langle \mathbf{v_1}, \mathbf{v_2} \rangle | \alpha_L, \gamma) \\ \propto & \left[\prod_{z=1}^{K} \mathrm{Dir}(\beta_z | \gamma) \right] \mathrm{Dir}(\pi_L | \alpha_L) \prod_{i=1}^{N_L} \pi_L^{\langle z_{i1}, z_{i2} \rangle} \beta_{z_{i1}}^{v_{i1}} \beta_{z_{i2}}^{v_{i2}} \end{aligned} \tag{6.1}$$

Since exact inference in the PSK model is intractable, we use a collapsed Gibbs sampler to perform approximate inference. A cluster pair for every link conditional on cluster pair assignments to all other links after collapsing π_L and $\boldsymbol{\beta}$, is sampled using the expression:

$$
\begin{aligned}
&p(z_i = \langle k_1, k_2\rangle | \langle v_{i1}, v_{i2}\rangle, \langle \mathbf{z_1}, \mathbf{z_2}\rangle^{\neg i}, \langle \mathbf{v_1}, \mathbf{v_2}\rangle^{\neg i}, \alpha_L, \gamma) \\
&= \frac{p(z_i = \langle k_1, k_2\rangle, \langle \mathbf{z_1}, \mathbf{z_2}\rangle^{\neg i}, \langle \mathbf{v_1}, \mathbf{v_2}\rangle | \alpha_L, \gamma)}{p(\langle \mathbf{z_1}, \mathbf{z_2}\rangle^{\neg i}, \langle \mathbf{v_1}, \mathbf{v_2}\rangle | \alpha_L, \gamma)} \\
&\propto \left(n^{L\neg i}_{\langle k_1,k_2\rangle} + \alpha_L\right) \frac{\left(n^{\neg i}_{k_1 v_{i1}} + \gamma\right)\left(n^{\neg i}_{k_2 v_{i2}} + \gamma\right)}{\left(\sum_v n^{\neg i}_{k_1 v} + |V|\gamma\right)\left(\sum_v n^{\neg i}_{k_2 v} + |V|\gamma\right)}
\end{aligned} \tag{6.2}
$$

The n's are counts of observations in the training set, where n_{kv} is the number of times a node v is observed under cluster k and $n^L_{\langle k_1,k_2\rangle}$ is the number of edges assigned to cluster pair $\langle k_1, k_2\rangle$. As before, counts with superscript $\neg i$ indicate that edge i is removed from the counts.

The cluster multinomial parameters and the cluster pair distributions of links are recovered using their point estimates after inference using the counts of observations:

$$
\beta^v_k = \frac{n_{kv} + \gamma}{\sum_{v'} n_{kv'} + |V|\gamma}\ , \quad \pi_L^{\langle k_1,k_2\rangle} = \frac{n^L_{\langle k_1,k_2\rangle} + \alpha_L}{\sum_{k'_1,k'_2} n^L_{\langle k'_1,k'_2\rangle} + K^2\alpha_L} \tag{6.3}
$$

A de-noised form of the entity-entity link matrix can also be recovered from the estimated parameters of the model. Let B be a matrix of dimensions $K \times |V|$ where row $k = \beta_k$, $k \in \{1, \cdots, K\}$. Let Z be a matrix of dimensions $K \times K$ s.t $Z_{p,q} = \pi_L^{\langle p,q\rangle}$. The de-noised matrix M of the strength of association between the nodes in V is given by $M = B^T Z B$.

6.3 Role Entropy Regularization

Each node $v \in V$ in the PSK model has a set of associated latent roles (z's) it plays when participating in edges. For every node v, we define a distribution $\mathbf{z}_v$ of dimension K where

$$
\tau^k_v = \sum_{v_{i1} \to v_{i2}} \frac{\mathbf{I}(v_{i1} = v)\mathbf{I}(z_{i1=k}) + \mathbf{I}(v_{i2} = v)\mathbf{I}(z_{i2} = k)}{\mathbf{I}(v_{i1} = v) + \mathbf{I}(v_{i2} = v)} \tag{6.4}
$$

where $\mathbf{I}(\cdot)$ takes the value 0 or 1 depending on the condition being true. The expression effectively computes $p(z = k|v)$, the latent role distribution of the node v. Figure 7.1 shows blue dashed edges from the latent role and edge-node variables to variables that represents the entropy of $\boldsymbol{\tau}_v$. Note that there is no distinction made between the occurrences of the node as a source or destination node, i.e., directionality of the edges is ignored while determining the latent role distribution.

As in the case of regularization in LDA-like models in earlier chapters, it should be noted that $\boldsymbol{\tau}_v$, the latent role distribution of a node, is not explicitly sampled during the generative process and is an aggregate function of the z and v variables that are generated. Since the different z and v values are independent draws conditioned on π_L and $\boldsymbol{\beta}$, any preference on a function that aggregates over different z and v values cannot be imposed by simply adjusting the parameters of the Dirichlet prior α_L.

We now introduce the role entropy regularization term by adding pseudo-observed variables, l_v, one for each node in V, which are noisy copies of $H(\boldsymbol{\tau}_v)$, to the generative process as seen in Fig. 7.1. $H(\boldsymbol{\tau}_v)$ is defined as $-\sum_k \tau_v^k \log_2 \tau_v^k$ and represents the Shannon entropy of $\boldsymbol{\tau}_v$. These pseudo-observed variables are drawn from Gaussians (truncated to limit mass between 0 and $\log_2 K$) with mean $H(\boldsymbol{\tau}_v)$ and variance $\sigma_{l_v}^2$ which is a hyperparameter to the model. The addition of the terms penalizes large entropies in the latent role distributions while retaining the generative nature of the model. The $\sigma_{l_v}^2$ parameter dictates the strictness of the penalty. The imposition of the penalty therefore allows us to overcome the independence assumption between the different z draws for a given node v.

The regularization term is defined as

$\prod_v l_v, \quad v \in V, l_v \sim \mathcal{N}\left(H(\boldsymbol{\tau}_v), \sigma_{l_v}^2\right)$. Therefore,

$$p\left(\prod_v l_v | \langle \mathbf{z_1}, \mathbf{z_2} \rangle, \langle \mathbf{v_1}, \mathbf{v_2} \rangle, \sigma_{l_v}^2\right) = \prod_v p(l_v | \langle \mathbf{z_1}, \mathbf{z_2} \rangle, \langle \mathbf{v_1}, \mathbf{v_2} \rangle, \sigma_{l_v}^2) \tag{6.5}$$

$$\propto \prod_v \exp \frac{-\left(l_v - H(\boldsymbol{\tau}_v)\right)^2}{2\sigma_{l_v}^2} \tag{6.6}$$

Similar to the case of word entropy regularization (section 3.2), since l_v is observed, i.e., its value is known during inference, the inference procedure tends to push the mean of the Gaussians i.e. $H(\boldsymbol{\tau}_v)$ close to the l_v values. We therefore set (*pseudo-observe*) l_v to 0 (or any desired small value) to coax the inference procedure to return low entropy latent role distributions for nodes. The variance parameter $\sigma_{l_v}^2$ can be used to adjust the tightness of the Gaussian to permit more or less entropy in the label distributions. In the limit, as $\sigma_{l_v}^2$ tends to 0, the model reduces to the stochastic block model since the regularization will require the entropies to be close to 0 implying that the distribution over latent roles has all its mass on one cluster. Similarly, as the variance

tends to ∞, the model reduces to a fully unconstrained mixed membership model. [1]

The joint distribution of the model with regularization is given by

$$\begin{aligned}\mathcal{L}^m &= p(\pi_L, \boldsymbol{\beta}, \mathbf{l_v}, \langle \mathbf{z_1}, \mathbf{z_2} \rangle, \langle \mathbf{v_1}, \mathbf{v_2} \rangle | \alpha_L, \gamma, \sigma_{l_v}^2) \\ &= \mathcal{L} \times \prod_v p(l_v | \langle \mathbf{z_1}, \mathbf{z_2} \rangle, \langle \mathbf{v_1}, \mathbf{v_2} \rangle, \sigma_{l_v}^2)\end{aligned} \tag{6.7}$$

Now, we derive the collapsed Gibbs sampling equation for the PSK model with role entropy.

$$\begin{aligned}&p(z_i = \langle k_1, k_2 \rangle | \mathbf{l_v}, \langle \mathbf{z_1}, \mathbf{z_2} \rangle^{\neg i}, \langle \mathbf{v_1}, \mathbf{v_2} \rangle | \alpha_L, \gamma, \sigma_{l_v}^2) \\ &= \frac{p(z_i = \langle k_1, k_2 \rangle, \langle \mathbf{z_1}, \mathbf{z_2} \rangle^{\neg i}, \langle \mathbf{v_1}, \mathbf{v_2} \rangle | \alpha_L, \gamma)}{p(\langle \mathbf{z_1}, \mathbf{z_2} \rangle^{\neg i}, \langle \mathbf{v_1}, \mathbf{v_2} \rangle | \alpha_L, \gamma)} \\ &\times \frac{p(\mathbf{l_v} | \langle \mathbf{z_1}, \mathbf{z_2} \rangle^{\neg i}, z_i = \langle k_1, k_2 \rangle, \langle \mathbf{v_1}, \mathbf{v_2} \rangle, \sigma_{l_v}^2)}{\sum_{k_1'} \sum_{k_2'} p(\mathbf{l_v} | \langle \mathbf{z_1}, \mathbf{z_2} \rangle^{\neg i}, z_i = \langle k_1', k_2' \rangle, \langle \mathbf{v_1}, \mathbf{v_2} \rangle, \sigma_{l_v}^2)}\end{aligned} \tag{6.8}$$

The first term in the product is the same as the unregularized model and we can replace the term with the expression from Equation 6.2. In the second term of the product, the denominator is not dependent on $\langle k_1, k_2 \rangle$ and can therefore be discarded as it only serves as a normalizing constant.

$$p(\mathbf{l_v} | \langle \mathbf{z_1}, \mathbf{z_2} \rangle, \langle \mathbf{v_1}, \mathbf{v_2} \rangle, \sigma_{l_v}^2) = \prod_v p(l_v | \langle \mathbf{z_1}, \mathbf{z_2} \rangle, \langle \mathbf{v_1}, \mathbf{v_2} \rangle, \sigma_{l_v}^2)$$

Terms in the product that do not pertain to v_{i1}, v_{i2}, z_{i1} and z_{i2} can be discarded since they are constants over all cluster pair label assignments. Therefore the second term in Equation 6.8 is only dependent on $\mathbf{z}_{v_{i1}}$ and $\mathbf{z}_{v_{i2}}$. The conditional distribution for collapsed Gibbs sampling equation can therefore be expressed as

$$\begin{aligned}&p(z_i = \langle k_1, k_2 \rangle | \mathbf{l_v}, \langle \mathbf{z_1}, \mathbf{z_2} \rangle^{\neg i}, \langle \mathbf{v_1}, \mathbf{v_2} \rangle, \alpha_L, \gamma, \sigma_{l_v}^2) \\ &\propto \left(n_{\langle k_1, k_2 \rangle}^{L \neg i} + \alpha_L \right) \frac{\left(n_{k_1 v_{i1}}^{\neg i} + \gamma \right) \left(n_{k_2 v_{i2}}^{\neg i} + \gamma \right)}{\left(\sum_v n_{k_1 v}^{\neg i} + |V|\gamma \right) \left(\sum_v n_{k_2 v}^{\neg i} + |V|\gamma \right)} \\ &\times \exp \frac{-\left(l_{v_{i1}} - H(\boldsymbol{\tau}_{v_{i1}}) \right)^2}{2\sigma_{l_v}^2} \exp \frac{-\left(l_{v_{i2}} - H(\boldsymbol{\tau}_{v_{i2}}) \right)^2}{2\sigma_{l_v}^2}\end{aligned} \tag{6.9}$$

[1] The pseudo-observed variables l_v can be modeled using a variety of distributions parameterized on $H(\boldsymbol{\tau}_v)$. As in the case of LDA regularization, we use Gaussian distributions because of its property of controllable peakiness (by adjusting the variance) around a desired mean and due to its minimal impact on sampling complexity.

where $\mathbf{z}_{v_{i1}}$ and $\mathbf{z}_{v_{i2}}$ use the assignment of $z_{i1} = k_1$ and $z_{i2} = k_2$.

It can be seen from the expression that adding the role entropy regularization is computationally inexpensive since the extra terms introduced in the collapsed Gibbs sampling expression only require the entropies of the current edge's source and destinate nodes' latent role distributions to be computed and does not require any computation over nodes and edges that do not participate in the edge being considered. An alternate way to achieve the same effect is to use a prior similar to ECD (Section 4.2) instead of regularization. As in the case of topic models, using such a prior, which is non-conjugate, complicates the inference procedure.

6.4 Cluster Volume Regularization

Cluster balance is an important aspect in clustering. Spectral clustering methods, which are relaxed versions of Ratio Cut and Normalized Cut [Shi and Malik, 2000], use different ways to define notions of cluster balance. In the case of Normalized Cut, the clusters are coaxed to have balanced volumes (which is defined as the sum of the degrees of the nodes in the cluster). To impose a similar preference in the PSK model, we propose a regularization scheme that prefers a higher entropy in the volume distribution $\mathbf{B}$, which is defined as:

$$B_k, k \in 1, \ldots K = \sum_{i=1}^{N_L} \frac{\mathbf{I}(z_{i1} = k) + \mathbf{I}(z_{i2} = k)}{2 * N_L} \tag{6.10}$$

The regularization term l_b (seen in Figure 7.1) is drawn from the Gaussian $\mathcal{N}(H(\mathbf{B}), \sigma_{l_b}^2)^{-1}$. It should be noted that the regularization term is the multiplicative inverse of the density.

Since the Gaussian term in the sampling equation below (Equation 6.11) is raised to the power -1, setting l_b to 0 will cause the sampling procedure to diminish the probability $p(l_b|H(\mathbf{B}), \sigma_{l_b}^2)$ by returning a mean for the Gaussian i.e. $H(\mathbf{B})$ that is far from 0, which means that $H(\mathbf{B})$ will tend to be high, implying that $\mathbf{B}$ will tend to be more balanced. The variance $\sigma_{l_b}^2$, like $\sigma_{l_v}^2$ in the case of role entropy regularization, controls the strictness of this preference. This value can be set to a lower value in cases where the network is believed to have more balanced clusters and can be set higher when bigger variations in the volumes of clusters is expected.

The joint distribution after adding volume and role entropy regularization terms to the PSK

Table 6.1: Dataset statistics.

Dataset	Nodes	Edges	Clusters	#Labels per-node
agblog	1222	33428	2	1
cora	2485	10138	7	1
citeseer	2114	7396	6	1
dolphin	62	318	2	1
football	115	1226	10	1
karate	34	156	2	1
polbooks	105	882	3	1
senate	98	9506	2	1
yeast	844	14780	15	2.5
blogcatalog	10,312	333,983	39	1.4
youtube	1,138,499	2,990,443	47	1.6

model is defined as $\mathcal{L}^{bm} = p(\pi_L, \boldsymbol{\beta}, l_b, \mathbf{l_v}, \langle \mathbf{z_1}, \mathbf{z_2} \rangle, \langle \mathbf{e_1}, \mathbf{e_2} \rangle | \alpha_L, \gamma, \sigma_{l_v}^2, \sigma_{l_b}^2)$ and is expressed as $\mathcal{L}^{bm} = \mathcal{L}^m \times p(l_b | \langle \mathbf{z_1}, \mathbf{z_2} \rangle, \sigma_{l_b}^2)$.

Similar to Equation 6.9, the collapsed Gibbs sampling equation for the latent cluster pair of an edge, with $\mathbf{B}$ using the assignment $z_{i1} = k_1$ and $z_{i2} = k_2$ is now defined as

$$
\begin{aligned}
& p(z_i = \langle k_1, k_2 \rangle | l_b, \mathbf{l_v}, \langle \mathbf{z_1}, \mathbf{z_2} \rangle^{\neg i}, \langle \mathbf{v_1}, \mathbf{v_2} \rangle, \alpha_L, \gamma, \sigma_{l_v}^2) \\
& \propto \left(n_{\langle k_1, k_2 \rangle}^{L \neg i} + \alpha_L \right) \frac{\left(n_{k_1 v_{i1}}^{\neg i} + \gamma \right) \left(n_{k_2 v_{i2}}^{\neg i} + \gamma \right)}{\left(\sum_v n_{k_1 v}^{\neg i} + |V|\gamma \right) \left(\sum_v n_{k_2 v}^{\neg i} + |V|\gamma \right)} \\
& \times \exp \frac{-(l_{v_{i1}} - H(\boldsymbol{\tau}_{v_{i1}}))^2}{2\sigma_{l_v}^2} \ \exp \frac{-(l_{v_{i2}} - H(\boldsymbol{\tau}_{v_{i2}}))^2}{2\sigma_{l_v}^2} \\
& \times \left(\exp \frac{-(l_b - H(\mathbf{B}))^2}{2\sigma_{l_b}^2} \right)^{-1}
\end{aligned}
\tag{6.11}
$$

6.5 Experimental Results

6.5.1 Datasets

We investigate the effects of regularization on a collection of datasets consisting of social networks, citation networks, yeast protein-protein interaction networks and other similar networks that have been studied in the sociology literature.

The first set of graphs [Balasubramanyan et al., 2010] are relatively small and have one known true cluster label for every node, which is used solely for evaluating the accuracy of node clustering. Statistics about the datasets are shown in the first 8 rows of Table 6.1.

The PolBooks dataset is a co-purchase network of 105 political books. Each book is labeled "liberal", "conservative", or "neutral", mostly in the first two categories. The Karate dataset describes the social network of friendships between 34 members of a karate club at a US university in the 1970. The Dolphin dataset is a social network of frequent associations between 62 dolphins in a community living off Doubtful Sound, New Zealand. The Football dataset represents instances of of American football games between Division IA colleges during the regular season in Fall 2000. [2]

The AGBlog [Adamic and Glance, 2005] dataset is from the domain of political blogs. The last two datasets, Citeseer and Cora [Lu and Getoor, 2003] link scientific publications. All of these datasets contain explicit links between nodes in the form of hyperlinks or citations. In constructing the graph from these datasets, we take the simplest approach possible; in each case the graph contains only undirected, unweighted edges. The Cora dataset contains papers from 7 categories and the CiteSeer dataset contains papers from 6 categories. The class names and class label distributions for these two datasets and the details of their construction is described in Lu and Getoor [2003]. Again, we extract the largest connected component from these datasets and end up with 2485 papers for the Cora dataset and 2114 papers for the CiteSeer dataset. An edge exist in the graph between node a and node b if paper a cites paper b or vice-versa.

Next, we study the Munich Institute for Protein Sequencing (MIPS) database which contains a

[2]The four datasets above were obtained from the UC Irvine Network Data Repository - http://networkdata.ics.uci.edu/

collection of protein interactions covering protein complex associations in yeast. We use a subset of this collection containing 844 proteins, for which all interactions were hand-curated. The proteins in the dataset are also annotated with functional categories based on the functions that they play. There are 15 top-level functional categories which are treated as known cluster labels. On average, a protein is annotated with 2.5 functional categories.

In addition to the smaller networks described above, we also run experiments on two larger benchmark networks, namely the BlogCatalog and YouTube datasets [Zafarani and Liu, 2009]. These larger networks also have known mixed-membership labels for nodes. On average across all nodes, nodes in the BlogCatalog dataset have 1.4 labels per node and nodes in the YouTube dataset have 1.6 labels per node. More statistics about these networks are in Table 6.1.

Experimental Setup

We evaluate the regularized and unregularized versions of the PSK model using the following metrics. The first metric used is average node entropy defined as $\sum_v H(\boldsymbol{\tau}_v)/|V|$. This metric shows the extent to which each node participates in multiple latent roles. The second metric used to evaluate the model is link perplexity which is a function of the likelihood of the edges in the dataset and is defined as

$$2^{-\frac{\sum_{v_{i1}\to v_{i2}} log_2 p(v_{i1}\to v_{i2})}{N_L}} \tag{6.12}$$

A lower perplexity value indicates an higher likelihood of data and a better fit.

For datasets with only one node per label, we can evaluate the model by checking its accuracy in predicting node labels. Nodes in the PSK model are associated with a distribution over clusters which can be obtained by normalizing β_z^v. Predicted class labels can be assigned to nodes using the 1-NN algorithm using the Jensen-Shannon distance between these cluster distributions as the metric to measure the distance between two nodes. We also evaluate the cluster labelling performance using NMI (Section 5.2).

Performance in the larger networks which have multiple labels per node is measured using micro and macro averaged F-1 measures of retrieving the known cluster labels. The prediction of multiple labels for a node is done in two stages. In the first stage, the Hungarian algorithm [Kuhn, 1955] is used to align true cluster labels to clusters in the model. Next, labels corresponding to elements

Table 6.2: Evaluation of regularization in the smaller datasets.

(all values computed using 10-fold cross-validation)

Dataset	Perplexity				Accuracy			
	Regularization				Regularization			
	None	Role	Volume	Both	None	Role	Volume	Both
agblog	2.47e+05	2.47e+0.5	2.33+05	**2.31+05**	0.921	0.925	0.922	**0.947**
citeseer	2.31e+06	1.73+06	1.60e+06	**1.51+06**	0.243	0.268	0.282	**0.291**
cora	3.41e+06	**2.27e+06**	2.52+06	2.62e+06	0.198	**0.268**	0.210	0.230
dolphin	2.10e+03	2.03e+03	2.03e+03	**2.00e+03**	0.871	**0.935**	0.897	0.881
football	1.31e+04	**0.79e+04**	0.96e+04	**0.79e+04**	0.161	**0.833**	0.515	0.560
karate	5.72e+02	**5.35e+02**	5.39e+02	5.54e+02	0.941	**1.00**	0.951	0.961
polbook	4.95e+03	4.91e+03	4.84e+03	**4.77e+03**	0.752	**0.778**	0.774	**0.778**
senatevote	3.50e+03	3.50e+03	**3.44e+03**	3.46e+03	0.969	**0.980**	**0.980**	0.971

from posterior role distributions of nodes that are above a threshold are treated as predicted labels.

For every dataset, we run experiments with 1) the baseline PSK model with no regularization, 2) PSK with role entropy regularization, 3) PSK with volume entropy and finally 4) PSK with both of the regularization terms. In all experiments, the number of clusters in the model is set to be the number of known clusters in the dataset. The collapsed Gibbs sampler is set to run for 100 iterations and the average of the last 10 samples is taken. Since collapsed Gibbs sampling results can vary depending on the random starting point, the accuracy and perplexity values reported are the means of 10 separate trials. The variance values $\sigma_{l_b}^2$ and $\sigma_{l_v}^2$ are set to 0.5 and we place priors which favor diagonal blocks over off-diagonal blocks by using a non-symmetric Dirichlet for α_L.

6.5.2 Results

First we study the direct impact of role entropy and volume entropy on the smaller networks measured by link perplexity and 1-NN clustering accuracy. It can be seen from Table 6.2 (bold values indicate the best performing model) that using role and volume entropy consistently decreases link perplexity and increases cluster prediction accuracy when compared to the baseline unregularized model. The improvements for both perplexity and accuracy for all variants of regularization is

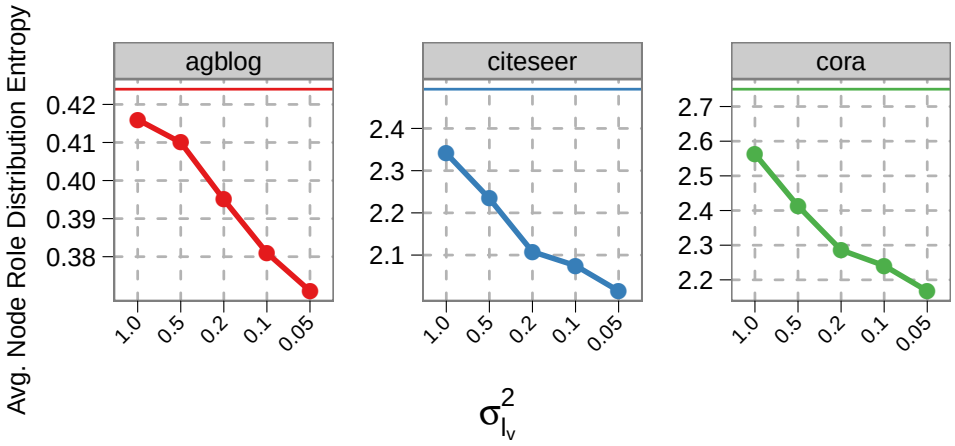


Figure 6.2: Effect of role distribution entropy (on 10% heldout tuning set)

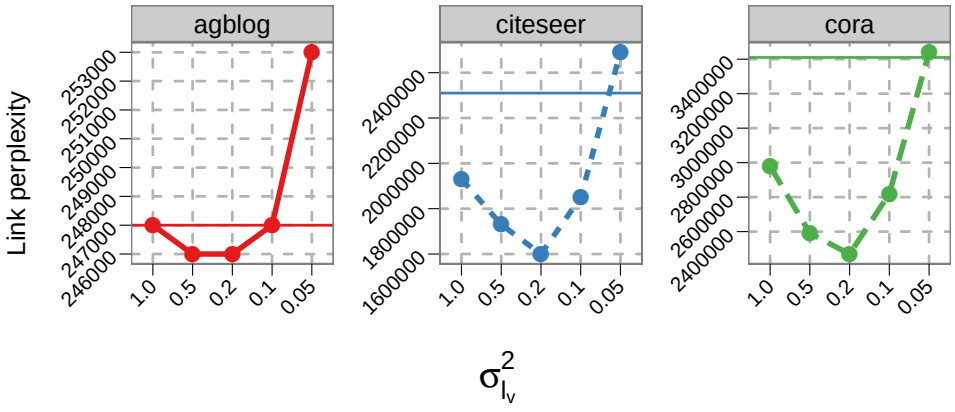


Figure 6.3: Perplexity with varying levels of Regularization (on 10% heldout tuning set)

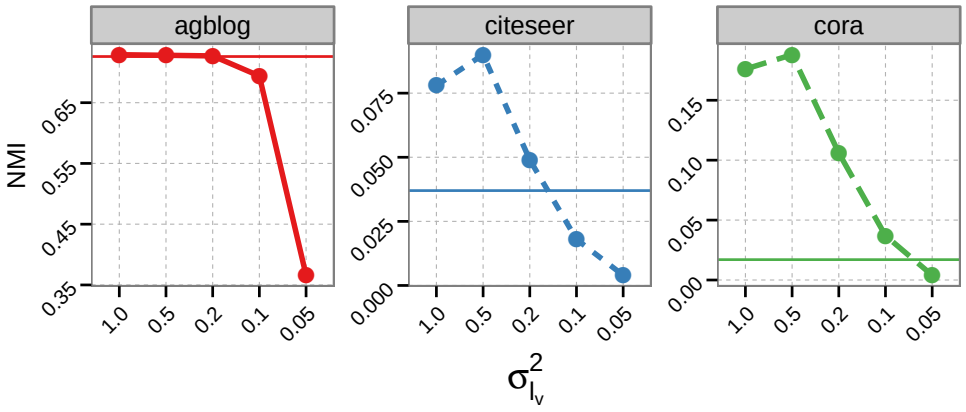


Figure 6.4: NMI with varying levels of Regularization (on 10% heldout tuning set)

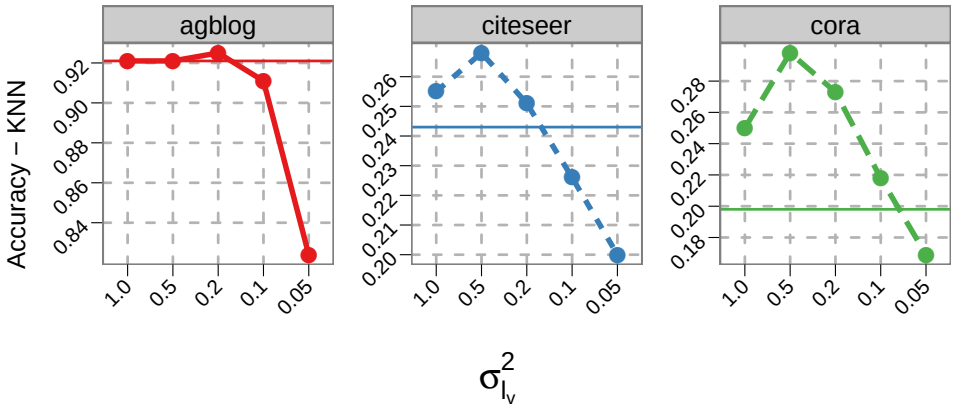


Figure 6.5: Cluster prediction accuracy with varying levels of Regularization (on 10% heldout tuning set)

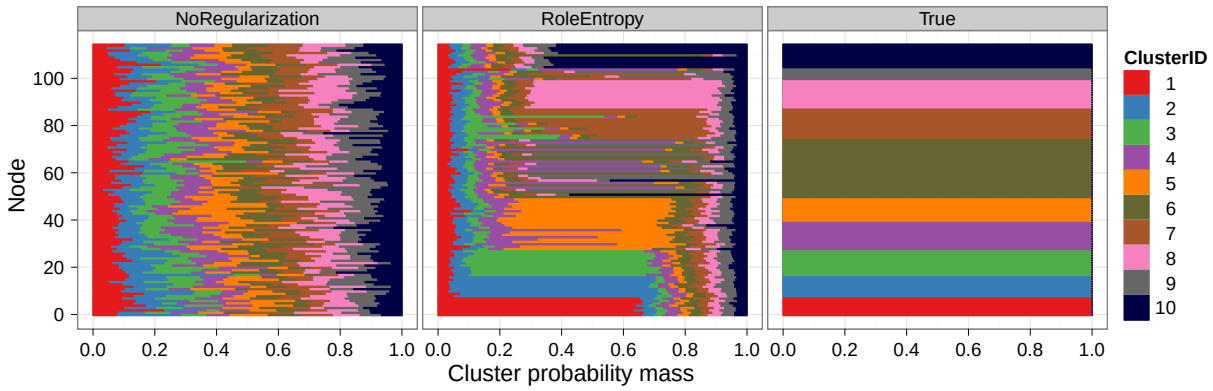


Figure 6.6: Role entropy demonstration: Inferred Latent Role distributions in the football network.

statistically significant at the 0.05 level using the Wilcoxon paired-sign test. The direct impact of role entropy regularization is illustrated further in Figure 6.2 which plots the average node entropy of 3 sample datasets obtained using different values of $\sigma_{l_v}^2$. It can be seen from the figure that the average node entropy in these datasets decreases as the variance parameter value is increased which shows that tightening the variance leads to models that tend closer to the stochastic block model where the entropy of the latent role distribution is 0.

Figure 6.6 shows the reduction of entropy in latent role distribution for one sample network more clearly. The figure shows a heatmap of the latent role distributions of each node in the network with nodes on the y-axis. The panel on the far right shows the known true label distributions with solid single colors in each row since nodes in the football dataset have one cluster label per node. The panel on the left shows the latent role distribution after inference using an unregularized PSK model. The middle panel shows the latent role distribution after inference with a role entropy regularized model. It can be seen clearly that the regularized model returns more peaky distributions with large probability masses residing in certain roles as compared to the unregularized model where the distributions are more equally distributed. The rows where the dominant color in the left and middle panels do not match the color in the right panel indicate cluster assignment errors.

Figures 6.3, 6.4 and 6.5 show how perplexity, NMI and clustering accuracy vary with different values for the variance term $\sigma_{l_v}^2$ on the same 3 sample networks used in Figure 6.2. The perplexity curves show a general U-shaped pattern that dips below the horizontal line representing the metric for the unregularized model, indicating a "sweet spot" for the variance value. Similarly in the

Table 6.3: Predicting cluster labels in mixed-membership datasets

Model	BlogCatalog			YouTube			Yeast		
	Micro F-1	Macro F-1	NMI	Micro F-1	Macro F-1	NMI	Micro F-1	Macro F-1	NMI
Unregularized	0.131	0.076	0.143	0.154	0.084	0.12	0.435	0.284	0.33
Role entropy	0.153*	0.077	0.151*	0.165*	0.086*	0.135*	0.485*	0.321*	0.352*
Volume entropy	0.154*	0.080*	0.170	0.171*	0.089*	0.132*	0.468*	0.305	0.346
Both	0.161*	0.082*	0.175*	0.184*	0.096	0.151	0.523*	0.310*	0.389*

All values computed using 10-fold cross-validation.

(* indicates a statistically significant improvement over the unregularized baseline).

3 accuracy plots, the regularized model accuracies rise above the baseline value with increasing variance values and then fall when it is increased further. This behavior is true for both measures internal to the model such as perplexity and external measures like NMI or clustering accuracy. At very low variance values, the model effectively approximates a single latent role stochastic block model since the nodes are restricted to only one role with high probability. This behavior tends to offer insufficient flexibility in modeling networks that inherently possess some mixed membership characteristics leading to drastic fall-offs in accuracies. These results indicate the although the smaller networks have only one true label per node, the actual structure of the networks does exhibit mixed-membership characteristics.

Finally, we evaluate the impact of regularization in the larger multi-labelled networks by checking the ability of the model to recover the known labels of nodes. Because nodes in these networks can have multiple labels, we use the micro and macro averaged F-1 measures to evaluate the clustering rather than accuracy. The models are also evaluated by computing NMI between the predicted and true known role distributiosn. Table 6.3 shows the F-1 measures and NMI values obtained from the clustering. It can be seen from the table that adding role and volume regularization improves performance in all 3 datasets. These networks have an average of 1.4 to 2.5 true labels per node and adding role entropy forces the model to restrict the number of roles a node can participate in which is a better fit to the true nature of the network. Volume entropy improves performance similarly by penalizing the formulation of trivially small clusters.

6.6 Related Work

The area of stochastic models of networks has been a fertile area of research; apart from MMSB and the PSK models, there have many various other approaches to modeling networks [Snijders, 2002]. Leskovec et al. [2010a] introduced Kronecker graphs, a framework that is flexible and can generate networks with commonly seen properties. Ho et al. [2012] argue that it is advantageous to represent networks as a set of triangular motifs rather than edges.

Airoldi et al. [2008] described a method to regularize the MMSB model to permit better fits for sparse graphs. The regularization techniques described here however, are designed to specifically influence the mixed-membership and balance characteristics of the network which is different from the goal of the regularization in the work described above. They also differ from previous regularization approaches through their use of pseudo-observed variables which allows the model to retain a generative story. The method presented in this chapter provides a general alternate way to impose preferences without the use of non-conjugate prior distributions by adding noisy copies of aggregate functions of latent variables to models which can be set to desired values to prefer desirable properties in the latent variable distributions.

6.7 Conclusion

We presented a general technique to impose preferences in latent variable models using the regularization framework described earlier in Chapter 3. We use the framework to regularize stochastic network models to control nodes' ability to take on different roles and to obtain balanced clusters. The regularization scheme permits the use of Gibbs sampling for inference with only an addition of a few terms to the sampling equations of the original PSK model. The technique of using pseudo-observed variables can also be used to impose other soft restrictions on networks such as controlling the incoming and outgoing latent roles separately and also to other stochastic network models such as MMSB. Experiments on real world network data both small and large, show that using slightly mixed-membership models using the regularization introduced provides better fits and consistently improves link perplexity and cluster label prediction. This chapter extends work presented earlier in Balasubramanyan et al. [2010].

Chapter 7

Joint Modeling of Network and Documents

7.1 Introduction

In previous chapters, we looked at enhancing LDA-like models and block models which share similarities with LDA. We presented different approaches to use partial prior knowledge and expectations about the latent structure while using mixed-membership models to uncover structure in data. These methods included enhancing the models to respect prior beliefs, and introducing techniques to include partially labeled data. In this chapter, we focus on situations where both document and interaction data are available. For instance in yeast biology, there exist both document data in the form of publications about the yeast organism and also interaction data in the form of protein-protein interaction networks. We present a model that combines these forms in information by jointly modeling documents and networks, in order to obtain a more informed structure from both components.

The task of modeling latent groups of entities from observed interactions is a commonly encountered problem. In social networks, for instance, we might want to identify sub-communities. In the biological domain, we might want to discover latent groups of proteins based on observed pairwise interactions. Mixed membership stochastic block models (MMSB) [Airoldi et al., 2008, Parkkinen et al., 2009] approach this problem by assuming that nodes in a graph represent entities

belonging to latent blocks with mixed membership, effectively capturing the notion that entities may arise from different sources and have different roles.

As discussed earlier, models like Latent Dirichlet Allocation(LDA) [Blei et al., 2003] treat text documents in a corpus as arising from mixtures of latent topics. Words in a document are potentially generated from different topics using topic specific word distributions. As described in Chapter 2 Link-LDA [Erosheva et al., 2004, Griffiths and Steyvers, 2004] additionally models other metadata in documents such as authors and entities, by treating a latent topic as a set of distributions, one for each metadata type. For instance, when modeling scientific publications from the biological domain, a latent topic could have a word distribution, an author distribution and a protein entity distribution.

In this chapter, we present a model, *Block-LDA*, that jointly generates text documents annotated with metadata about associated entities and external links between pairs of entities allowing it to use supplementary annotated text to influence and improve link modeling. The text documents are modeled as bags of entities of different types and the network is modeled as edges between entities of a source type to a destination type. Consider an example of a corpus of publications about the yeast organism and a network of protein-protein interactions in yeast. These publications are further annotated by experts with lists of proteins that are discussed in them. Therefore each publication could be modeled as a collection of bags *vis-a-vis* bag of body-words, bag of authors, bag of proteins discussed in the paper, etc. Similarly, the network could be a collection of protein-protein interactions independently observed.

The model merges the idea of latent topics in topic models with blocks in stochastic block models. The joint modeling permits sharing of information about the latent topics between the network structure and text, resulting in more coherent topics. Co-occurrence patterns in entities and words related to them aid the modeling of links in the graph. Likewise, entity-entity links provide provide clues about topics in the text. We also propose a method to perform approximate inference in the model using a collapsed Gibbs sampler, since exact inference in the joint model is intractable.

We then use the model to organize a large collection of literature about yeast biology to enable topic oriented browsing and retrieval from the literature. The analysis is performed using the mixed

membership topic modeling to uncover latent structure in document corpora by identifying broad topics that are discussed in it. This approach complements traditional information retrieval tasks where the objective is to fulfill very specific information needs. By using joint modeling, we are able to use other sources of domain information related to the domain in addition to literature, In the case of yeast biology, an example of such a resource is a database of known protein-protein interactions (PPI) which have been identified using wetlab experiments. We perform data fusion by combining text information from articles and the database of yeast protein-protein interactions, by using a latent variable model — Block-LDA [Balasubramanyan and Cohen, 2011] that jointly models the literature and PPI networks.

We evaluate the ability of the topic models to return meaningful topics by inspecting the top papers and proteins that pertain to them. We compare the performance of the joint model, i.e., Block-LDA with a model that only considers the text corpora by asking a yeast biologist to evaluate the coherence of topics and the relevance of the retrieved articles and proteins. This evaluation serves to test the utility of Block-LDA on a real task as opposed to an internal evaluation (such as by using perplexity metrics for example). Our evaluaton shows that the joint model outperforms the text-only approach both in topic coherence and in top paper and protein retrieval as measured by precision@10 values.

7.2 Block-LDA

The Block-LDA model (plate diagram in Figure 7.1) enables sharing of information between the component on the left that models links between pairs of entities represented as edges in a graph with a block structure, and the component on the right that models text documents, through shared latent topics. More specifically, the distribution over the entities of the type that are linked is shared between the block model and the text model.

The component on the right uses Link-LDA (Chapter 2) to model documents as sets of "bags of entities", each bag corresponding to a particular type of entity. Every entity type has a topic wise multinomial distribution over the set of entities that can occur as an instance of the entity type.

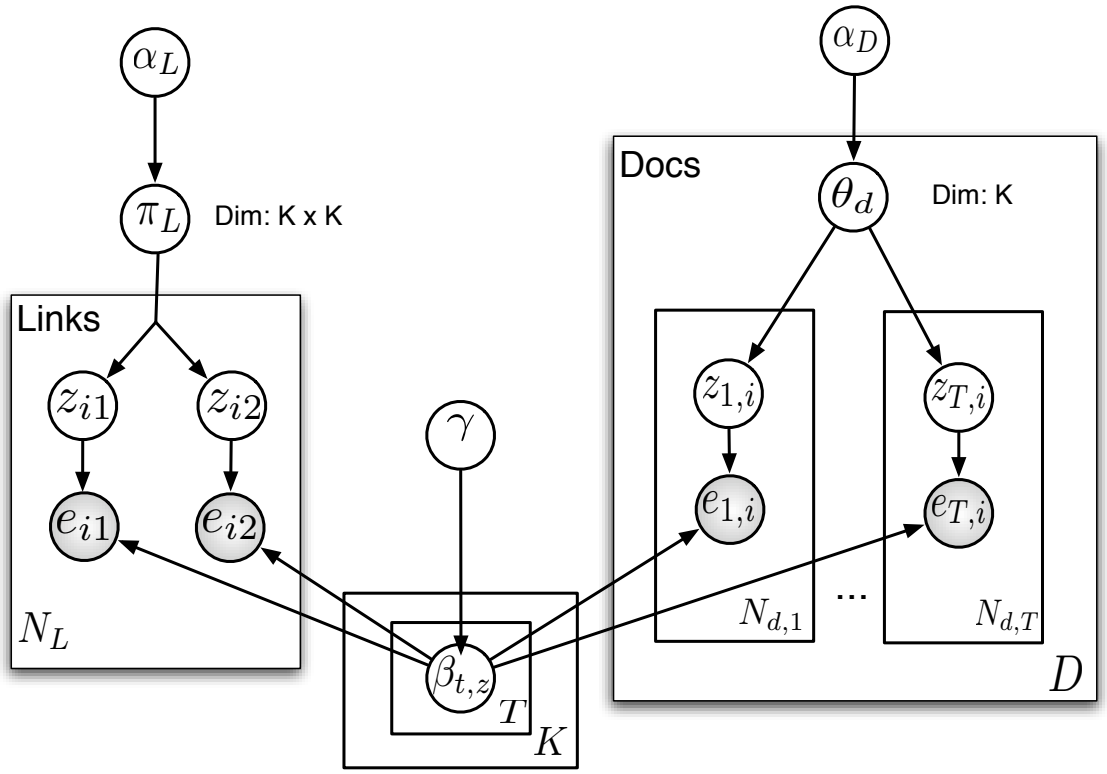


Figure 7.1: Block-LDA: plate diagram

The component on the left in the figure is a generative model (figure 6.1) for graphs representing entity-entity links with an underlying block structure, derived from the PSK introduced by Parkkinen et al. [2009], as described earlier in section 6.2.

Let K be the number of latent topics(blocks) we wish to recover. Assuming documents consist of T different types of entities (i.e., each document contains T bags of entities), and that links in the graph are between entities of type t_l and t_r, the generative process is as follows.

1. Generate topics:
   - For each type $t \in 1, \ldots, T$, and topic $z \in 1, \ldots, K$, sample $\beta_{t,z} \sim$ Dirichlet(γ), the topic specific entity distribution.
2. Generate documents. For every document $d \in \{1 \ldots D\}$:
   - Sample $\theta_d \sim$ Dirichlet(α_D) where θ_d is the topic mixing distribution for the document.
   - For each type t and its associated set of entity mentions $e_{t,i}$, $i \in \{1, \cdots, N_{d,t}\}$:

K - the number of topics (therefore resulting in K^2 blocks in the network)

α_L - Dirichlet prior for the topic pair distribution for links

α_D - Dirichlet prior for document specific topic distributions

γ - Dirichlet prior for topic multinomials

π_L - multinomial distribution over topic pairs for links

θ_d - multinomial distribution over topics for document d

T - the number of types of entities in the corpus

$\beta_{t,z}$ - multinomial over entities of type t for topic z

D - number of documents in the corpus

$z_{t,i}$ - topic chosen for the i-th entity of type t in a document

$e_{t,i}$ - the i-th entity of type t occurring in a document

N_L - number of links in the network

z_{i1} and z_{i2} - topics chosen for the two nodes participating in the i-th link

e_{i1} and e_{i2} - the two nodes participating in the i-th link

Figure 7.2: Variables in Block-LDA

- Sample a topic $z_{t,i} \sim$ Multinomial(θ_d)
- Sample an entity $e_{t,i} \sim$ Multinomial$(\beta_{t,z_{t,i}})$

3. Generate the link matrix of entities of type t_l:

- Sample $\pi_L \sim$ Dirichlet(α_L) where π_L describes a distribution over the Cartesian product of topics, for links in the dataset.
- For every link $e_{i1} \rightarrow e_{i2}$, $i \in \{1 \cdots N_L\}$:
  - Sample a topic pair $\langle z_{i1}, z_{i2} \rangle \sim$ Multinomial(π_L)
  - Sample $e_{i1} \sim$ Multinomial$(\beta_{t_l,z_{i1}})$
  - Sample $e_{i2} \sim$ Multinomial$(\beta_{t_r,z_{i2}})$

Note that unlike the MMSB model introduced by Airoldi et al. [2008], this model generates only realized links between entities.

Given the hyperparameters α_D, α_L and γ, the joint distribution over the documents, links, their topic distributions and topic assignments is given by

$$
\begin{aligned}
&p(\pi_L, \boldsymbol{\theta}, \boldsymbol{\beta}, \mathbf{z}, \mathbf{e}, \langle \mathbf{z_1}, \mathbf{z_2} \rangle, \langle \mathbf{e_1}, \mathbf{e_2} \rangle | \alpha_D, \alpha_L, \gamma) \propto \\
&\prod_{z=1}^{K} \prod_{t=1}^{T} \mathrm{Dir}(\beta_{t,z} | \gamma_t) \times \\
&\prod_{d=1}^{D} \mathrm{Dir}(\theta_d | \alpha_D) \prod_{t=1}^{T} \prod_{i=1}^{N_{d,t}} \theta_d^{z_{t,i}^{(d)}} \beta_{t, z_{t,i}^{(d)}}^{e_{t,i}} \times \\
&\mathrm{Dir}(\pi_L | \alpha_L) \prod_{i=1}^{N_L} \pi_L^{\langle z_{i1}, z_{i2} \rangle} \beta_{t_l, z1}^{e_{i1}} \beta_{t_r, z2}^{e_{i2}}
\end{aligned}
\tag{7.1}
$$

A commonly required operation when using models like Block-LDA is to perform inference on the model to query the topic distributions and the topic assignments of documents and links. Due to the intractability of exact inference in the Block-LDA model, a collapsed Gibbs sampler is used to perform approximate inference, just as in the case of Link-LDA. The sampler samples a latent topic for an entity mention of type t in the text corpus conditioned on the assignments to all other entity mentions using the following expression (after collapsing θ_D, and similar to Equation 2.2):

$$
p(z_{t,i} = z | e_{t,i}, \mathbf{z}^{\neg \mathbf{i}}, \mathbf{e}^{\neg \mathbf{i}}, \alpha_D, \gamma) \propto (n_{dz}^{\neg i} + \alpha_D) \frac{n_{zte_{t,i}}^{\neg i} + \gamma}{\sum_{e'} n_{zte'}^{\neg i} + |V_t| \gamma} \tag{7.2}
$$

Similarly, we sample a topic pair for every link conditional on topic pair assignments to all other links after collapsing π_L using the expression:

$$
\begin{aligned}
&p(\mathbf{z_i} = \langle z_1, z_2 \rangle | \langle e_{i1}, e_{i2} \rangle, \mathbf{z}^{\neg \mathbf{i}}, \langle \mathbf{e_1}, \mathbf{e_2} \rangle^{\neg \mathbf{i}}, \alpha_L, \gamma) \\
&\propto \left(n_{\langle z_1, z_2 \rangle}^{L \neg i} + \alpha_L \right) \frac{\left(n_{z_1 e_{i1} t_l}^{\neg i} + \gamma \right) \left(n_{z_2 e_{i2} t_r}^{\neg i} + \gamma \right)}{\left(\sum_e n_{z_1 e t_l}^{\neg i} + |V_{t_l}| \gamma \right) \left(\sum_e n_{z_2 e t_r}^{\neg i} + |V_{t_r}| \gamma \right)}
\end{aligned}
\tag{7.3}
$$

V_t refers to the set of all entities of type t. The n's refer to number of topic assignments in the data.

- n_{zet} - the number of times an entity e of type t is observed under topic z
- n_{dz} - the number of entities (of any type) with topic z in document d
- $n_{\langle z_1, z_2 \rangle}^{L}$ - count of links assigned to topic pair $\langle z_1, z_2 \rangle$

The topic multinomial parameters and the topic distributions of links and documents are easily recovered using their MAP estimates after inference using the counts of observations.

$$\beta_{t,z}^{(e)} = \frac{n_{zet} + \gamma}{\sum_{e'} n_{ze't} + |E_t|\gamma}, \tag{7.4}$$

$$\theta_d^{(z)} = \frac{n_{dz} + \alpha_D}{\sum_{z'} n_{dz'} + K\alpha_D} \text{and} \tag{7.5}$$

$$\pi_L^{\langle z_1, z_2 \rangle} = \frac{n_{\langle z_1, z_2 \rangle} + \alpha_L}{\sum_{z_1', z_2'} n_{\langle z_1', z_2' \rangle} + K^2\alpha_L} \tag{7.6}$$

A de-noised form of the entity-entity link matrix can also be recovered from the estimated parameters of the model. Let B_t be a matrix of dimensions $K \times |E_t|$ where row $k = \beta_{t,k}$, $k \in \{1, \cdots, K\}$. Let Z be a matrix of dimensions $K \times K$ s.t $Z_{p,q} = \sum_{i=1}^{N_L} \mathbf{I}(z_{i1} = p, z_{i2} = q)$. The de-noised matrix M of the strength of association between the entities in E_{t_l} is given by $M = B_{t_l}^T Z B_{t_r}$.

7.3 Datasets

The Munich Institute for Protein Sequencing (MIPS) database [Mewes et al., 2004] includes a hand-crafted collection of protein interactions covering 8000 protein complex associations in yeast. We use a subset of this collection containing 844 proteins, for which all interactions were hand-curated (Figure 7.3(a)). The MIPS institute also provides a set of functional annotations for each protein which are organized in a tree, with 15 nodes at the first level (shown in Table 7.1). The 844 proteins participating in interactions are mapped to these 15 functional categories with an average of 2.5 annotations per protein.

We also use another dataset of protein-protein interactions in yeast that were observed as a result of wetlab experiments by our collaborators in John Woolford's lab at the Department of Biology at Carnegie Mellon University. This dataset consists of 635 interactions that deal primarily with ribosomal proteins and assembly factors in yeast.

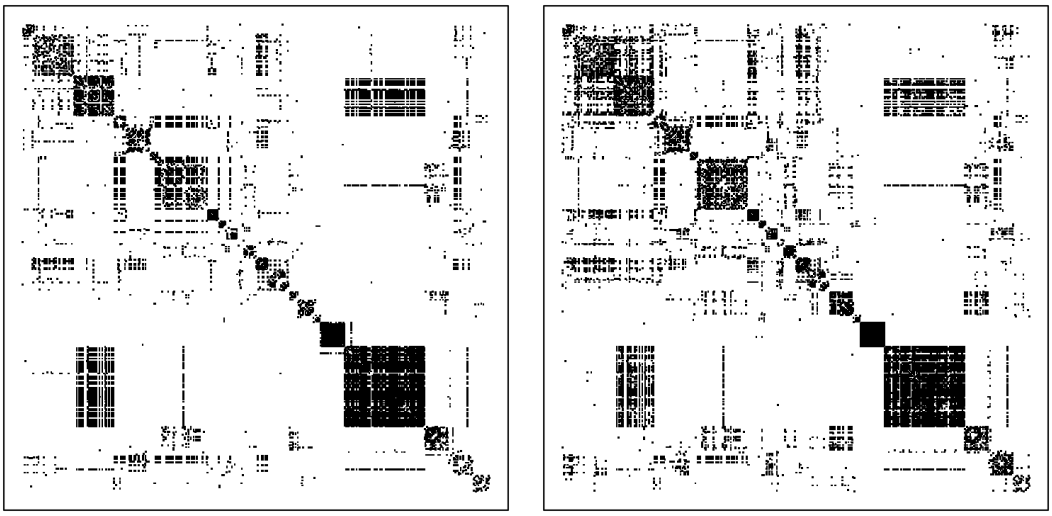
In addition to the MIPS PPI data, we use a text corpus that is derived from the repository of scientific publications at PubMed®. PubMed is a free, open-access on-line archive of over 18 million biological abstracts and bibliographies, including citation lists, for papers published since

Metabolism
Cellular communication/signal transduction mechanism
Cell rescue, defense and virulence
Regulation of / interaction with cellular environment
Cell fate
Energy
Control of cellular organization
Cell cycle and DNA processing
Subcellular localisation
Transcription
Protein synthesis
Protein activity regulation
Transport facilitation
Protein fate (folding, modification, destination)
Cellular transport and transport mechanisms

Table 7.1: List of functional categories

1948 (U.S. National Library of Medicine 2008). The subset we work with consists of approximately 40,000 publications about the yeast organism that have been curated in the Saccharomyces Genome Database (SGD) [Dwight et al., 2004] with annotations of proteins that are discussed in the publication. We further restrict the dataset to only those documents that are annotated with at least one protein from the MIPS database. This results in a MIPS-protein annotated document collection of 15,776 publications. The publications in this set were written by a total of 47,215 authors. We tokenize the titles and abstracts based on white space, lowercase all tokens and eliminate stopwords. Low frequency (< 5 occurrences) terms are also eliminated. The vocabulary contains 45,648 words.

To investigate the co-occurrence patterns of proteins annotated in the abstracts, we construct a co-occurrence matrix. From every abstract, a link is constructed for every pair of annotated protein mentions. Additionally, protein mentions that occur fewer than 5 times in the corpus are discarded.



(a) MIPS interactions (b) Co-occurences in text

Figure 7.3: Observed protein-protein interactions compared to thresholded co-occurrence in text

Figure 7.3(b) shows the resultant matrix, which looks very similar to the MIPS PPI matrix in Figure 7.3(a). This suggests that joint modeling of the protein annotated text with the PPI information has the potential to be beneficial. The nodes representing proteins in 7.3(b) and 7.3(a) are ordered by their cluster ids, obtained by clustering them using k-means clustering, treating proteins as 15-bit vectors of functional category annotations.

The Enron email corpus[Shetty and Adibi, 2004] is a large publicly available collection of email messages subpoenaed as part of the investigation by the Federal Energy Regulatory Commission (FERC). The dataset contains 517,437 messages in total. Although the Enron Email Dataset contains the email folders of 150 people, two people appear twice with different usernames, and one user's emails consist solely of automated emails, resulting in 147 unique people in the dataset. For the text component of the model, we use all the emails in the Sent[1] folders of the 147 users' mailboxes, resulting in a corpus of 96,103 emails. Messages are annotated with mentions of people from the set of 147 Enron employees if they are senders or recipients of the email. Mentions of people outside of the 147 persons considered are dropped. While extracting text from the email messages, "quoted" messages are eliminated using a heuristic which looks for a "Forwarded message" or

[1] "sent", "sent_items" and "_sent_mail" folders in users' mailboxes were treated as "Sent" folders

"Original message" delimiter. In addition, lines starting with a ">" are also eliminated. The emails are then tokenized after lowercasing the entire message, using whitespace and punctuation marks as word delimiters. Words occurring fewer than 5 times in the corpus are discarded. The vocabulary of the corpus consists of 32,880 words.

For the entity links component of the model, we build an email communication network by constructing a link between the sender and every recipient of an email message, for every email in the corpus. Recipients of the emails include people directly addressed in the "TO" field and people included in the "CC" and "BCC" fields. Similar to the text component, only links between the 147 Enron employees are considered. The link dataset generated in this manner has 200,404 links. Figure 7.8(a) shows the email network structure. The nodes in the matrix representing people are ordered by cluster ids obtained by running k-means clustering on the 147 people. Each person s is represented by a vector of length 147, where the elements in the vector are normalized counts of the number of times an email is sent by s to the person indicated by the element.

7.4 Experimental Results

We present results from experiments using Block-LDA to model the Yeast and Enron datasets described in Section 7.3.

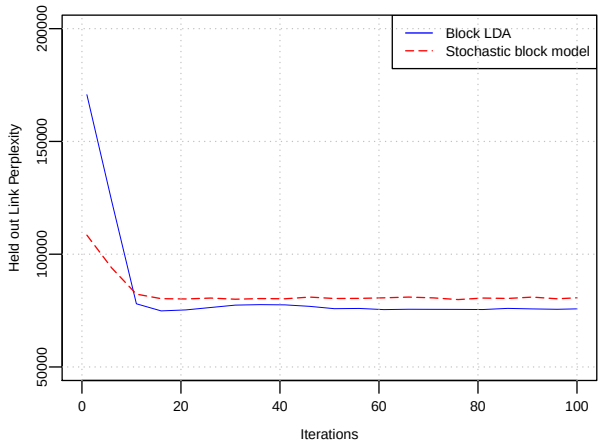
7.4.1 Results from the Yeast dataset

Perplexity and convergence

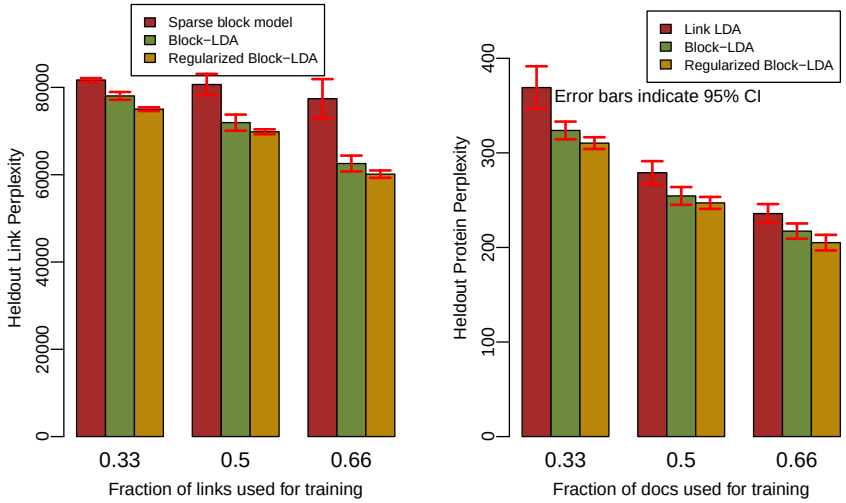
First, we investigate the convergence properties of the collapsed Gibbs sampler used for inference in Block-LDA by observing link perplexity on heldout data at different epochs. Link perplexity of set of links L is defined as

$$\exp\left(\frac{\sum_{e_1 \to e_2 \in L} \log\left(\sum_{\langle z_1, z_2\rangle} \pi^{\langle z_1, z_2\rangle} \beta^{(e_1)}_{t_l, z_1} \beta^{(e_1)}_{t_l, z_2}\right)}{|L|}\right) \tag{7.7}$$

Figure 7.4(a) shows the convergence of the link perplexity using Block LDA and a baseline non-mixed-membership block model on the PPI+SGD dataset with 20% of the full dataset heldout



(a) Collapsed Gibbs sampler convergence

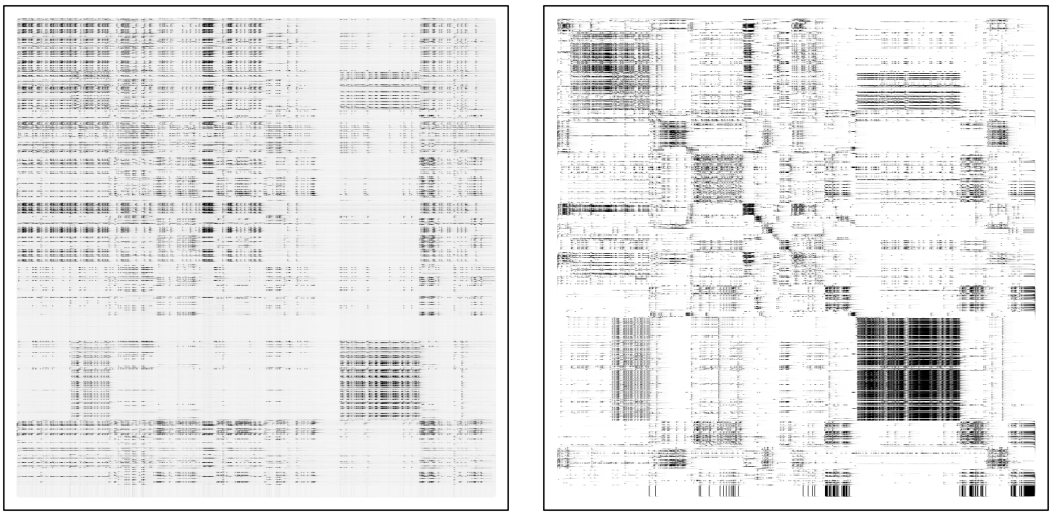


(b) Gain in perplexity through joint modeling

Figure 7.4: Perplexity in the MIPS PPI+SGD dataset (perplexity computed using 10-fold cross-validation)

for testing. The number of topics K is set at 15 since our aim is to recover topics that can be aligned with the 15 protein functional categories. α_D and α_L are set to 0.2. It can be observed that the collapsed Gibbs sampler burns-in after about 20 iterations.

Next, we perform two sets of experiments with the PPI+PubMed Central dataset. The text data has 3 types of entities in each document - words, authors and protein annotations with the



(a) Sparse block model (b) Block-LDA

Figure 7.5: Inferred protein-protein interactions

PPI data linking proteins. In the first set of experiments, we evaluate the model using perplexity of heldout protein-protein interactions using increasing amounts of the PPI data for training.

All the 15,773 documents in the SGD dataset are used when textual information is used. When text is not used, the model is equivalent to using only the left half of Figure 7.1. Figures 7.5(a) and 7.5(b) shows the posterior likelihood of protein-protein interactions recovered using the sparse block model and using Block-LDA respectively. In the other set of experiments, we evaluate the model using protein perplexity in heldout text using progressively increasing amounts of text as training data. All the links in the PPI dataset are used in these experiments when link data is used. When link data are not used, the model reduces to Link LDA. For the regularized Block-LDA experiments, the freedom of proteins to span multiple latent roles in the text corpus is restricted using the regularization technique presented in Section 3.2. The variance hyperparameter $\alpha_{l_w}^2$ is set to 0.5 following experimental results from the previous chapter. In all experiments, the collapsed Gibbs sampler is run until the held out perplexity stabilizes to a nearly constant value ($\approx$ 80 iterations).

Figure 7.4(b) shows the gains in perplexity in the two sets of experiments with different amounts of training data. The perplexity values are averaged over 10 trials. In both sets of experiments, it can be seen that Block-LDA results in lower perplexities than using links/text alone. Regularizing

the model by restricting proteins' freedom to span latent roles provides a further improvement. These results indicate that co-occurrence patterns of proteins in text contain information about protein interactions which Block-LDA is able to utilize through joint modeling. Our conjecture is that the protein co-occurrence information in text is a noisy approximation of the PPI data.

Table 7.2 shows the top words, proteins and authors for sample topics induced by running Block-LDA over the full PPI+SGD dataset. These topics provide a qualitative feel for the structure hat emerge using the model. The collapsed Gibbs sampling procedure was run until the perplexity value stabilized (around 80 iterations) and the number of topics was set to 15. The topic tables were then analyzed and a title and an analysis of the topic added, after the inference procedure. Details about proteins and yeast researchers were obtained on the SGD [2] website to understand the function of the top proteins in each topic and to get an idea of the research profile of the top authors mentioned.

Topic Coherence

An useful application of latent block modeling approaches is to understand the underlying nature of data. We conduct three different evaluations of the emergent topics. Firstly, we obtain topics from only the text corpus using a model that comprises of the right half of Figure 7.1 which is equivalent to using the Link-LDA model. For the second evaluation, we use the Block-LDA model that is trained on the text corpus and the MIPS protein-protein interaction database. Finally, for the third evaluation, we replace the MIPS PPI database with the interaction obtained from the wetlab experiments. In all the cases, we set K, the number of topics to be 15. In each variant, we represent documents as 3 sets of entities i.e. the words in the abstracts of the article, the set of proteins associated with the article as indicated in the SGD database and finally the authors who wrote the article. Each topic therefore consists of 3 different multinomial distributions over the sets of the 3 kinds of entities described.

Topics that emerge from the different variants can possibly be assigned different indices even when they discuss the same semantic concept. To compare topics across variants, we need a method to determine which topic indices from the different variants correspond to the same se-

[2] http://www.yeastgenome.org

Words	mutant, mutants, gene, cerevisiae, growth, type, mutations, saccharomyces, wild, mutation, strains, strain, phenotype, genes, deletion
Proteins	rpl20b, rpl5, rpl16a, rps5, rpl39, rpl18a, rpl27b, rps3, rpl23a, rpl1b, rpl32, rpl17b, rpl35a, rpl26b, rpl31a
Authors	klis_fm, bussey_h, miyakawa_t, toh-e_a, heitman_j, perfect_jr, ohya_y_ws, sherman_f, latge_jp, schaffrath_r, duran_a, sa-correia_i, liu_h, subik_j, kikuchi_a, chen_j, goffeau_a, tanaka_k, kuchler_k, calderone_r, nombela_c, popolo_l, jablonowski_d, kim_j
Analysis	A common experimental procedure is to induce random mutations in the "wild-type" strain of a model organism (e.g., saccharomyces cerevisiae) and then screen the mutants for interesting observable characteristics (i.e. phenotype). Often the phenotype shows slower growth rates under certain conditions (e.g. lack of some nutrient). The RPL* proteins are all part of the larger (60S) subunit of the ribosome. The first two biologists, Klis and Bussey's research use this method.

(a) Analysis of Mutations

Words	binding, domain, terminal, structure, site, residues, domains, interaction, region, subunit, alpha, amino, structural, conserved, atp
Proteins	rps19b, rps24b, rps3, rps20, rps4a, rps11a, rps2, rps8a, rps10b, rps6a, rps10a, rps19a, rps12, rps9b, rps28a
Authors	naider_f, becker_jm, leulliot_n, van_tilbeurgh_h, melki_r, velours_j, graille_m_s, janin_j, zhou_cz, blondeau_k, ballesta_jp, yokoyama_s, bousset_l, vershon_ak, bowler_be, zhang_y, arshava_b, buchner_j, wickner_rb, steven_ac, wang_y, zhang_m, forgac_m, brethes_d
Analysis	Protein structure is an important area of study. Proteins are composed of amino-acid residues, functionally important protein regions are called domains, and functionally important sites are often "converved" (i.e., many related proteins have the same amino-acid at the site). The RPS* proteins all part of the smaller (40S) subunit of the ribosome. Naider, Becker, and Leulliot study protein structure.

(b) Protein Structure

Words	transcription, ii, histone, chromatin, complex, polymerase, transcriptional, rna, promoter, binding, dna, silencing, h3, factor, genes
Proteins	rpl16b, rpl26b, rpl24a, rpl18b, rpl18a, rpl12b, rpl6b, rpp2b, rpl15b, rpl9b, rpl40b, rpp2a, rpl20b, rpl14a, rpp0
Authors	workman_jl, struhl_k, winston_f, buratowski_s, tempst_p, erdjument-bromage_h, kornberg_rd_a, svejstrup_jq, peterson_cl, berger_sl, grunstein_m, stillman_dj, cote_j, cairns_br, shilatifard_a, hampsey_m, allis_cd, young_ra, thuriaux_p, zhang_z, sternglanz_r, krogan_nj, weil_pa, pillus_l
Analysis	In transcription, DNA is unwound from histone complexes (where it is stored compactly) and converted to RNA. This process is controlled by transcription factors, which are proteins that bind to regions of DNA called promoters. The RPL* proteins are part of the larger subunit of the ribosome, and the RPP proteins are part of the ribosome stalk. Many of these proteins bind to RNA. Workman, Struhl, and Winston study transcription regulation and the interaction of transcription with the restructuring of chromatin (a combination of DNA, histones, and other proteins that comprises chomosomes).

(c) Chromosome remodeling and transcription

Words	rna, mrna, nuclear, translation, pre, ribosomal, processing, complex, rrna, export, splicing, factor, required, prion, binding
Proteins	sup35, rpl3, rps2, rpl18a, rpl6a, rpl7a, rpl42b, rpl5, rpl18b, rps0b, rpl22a, rps11b, rpl27b, rpl32, rpl7b
Authors	tollervey_d, hurt_e, parker_r, wickner_rb, seraphin_b, corbett_ah, silver_pa, hinnebusch_c, baserga_sj, rosbash_m, beggs_jd, jacobson_a, liebman_sw, linder_p, petfalski_e, luhrmann_r, fromont-racine_m, ter-avanesyan_md, johnson_aw, raue_ha, keller_w, schwer_b, wente_sr, tuite_mf
Analysis	Translation is conversion of DNA to mRNA, a process that is followed by splicing (in which parts of the mRNA are removed). sup35 is a protein that terminates transcription; it also exists as a misfolded protein called a "prion". Tollervey, Hurt, and Parker study RNA processing and export.

(d) RNA maturation

Table 7.2: Top words, proteins and authors: topics obtained using Block-LDA on the PPI+SGD dataset

Variant	Num. Coherent Topics
Only Text	12 / 15
Text + MIPS	13 / 15
Text + Wetlab	15 / 15

Table 7.3: Topic Coherence Evaluation

mantic concept. To obtain the mapping between topics from each variant, we use the Hungarian algorithm [Kuhn, 1955] to solve the assignment problem, where the cost of aligning topics together is determined using the Jensen-Shannon divergence measure.

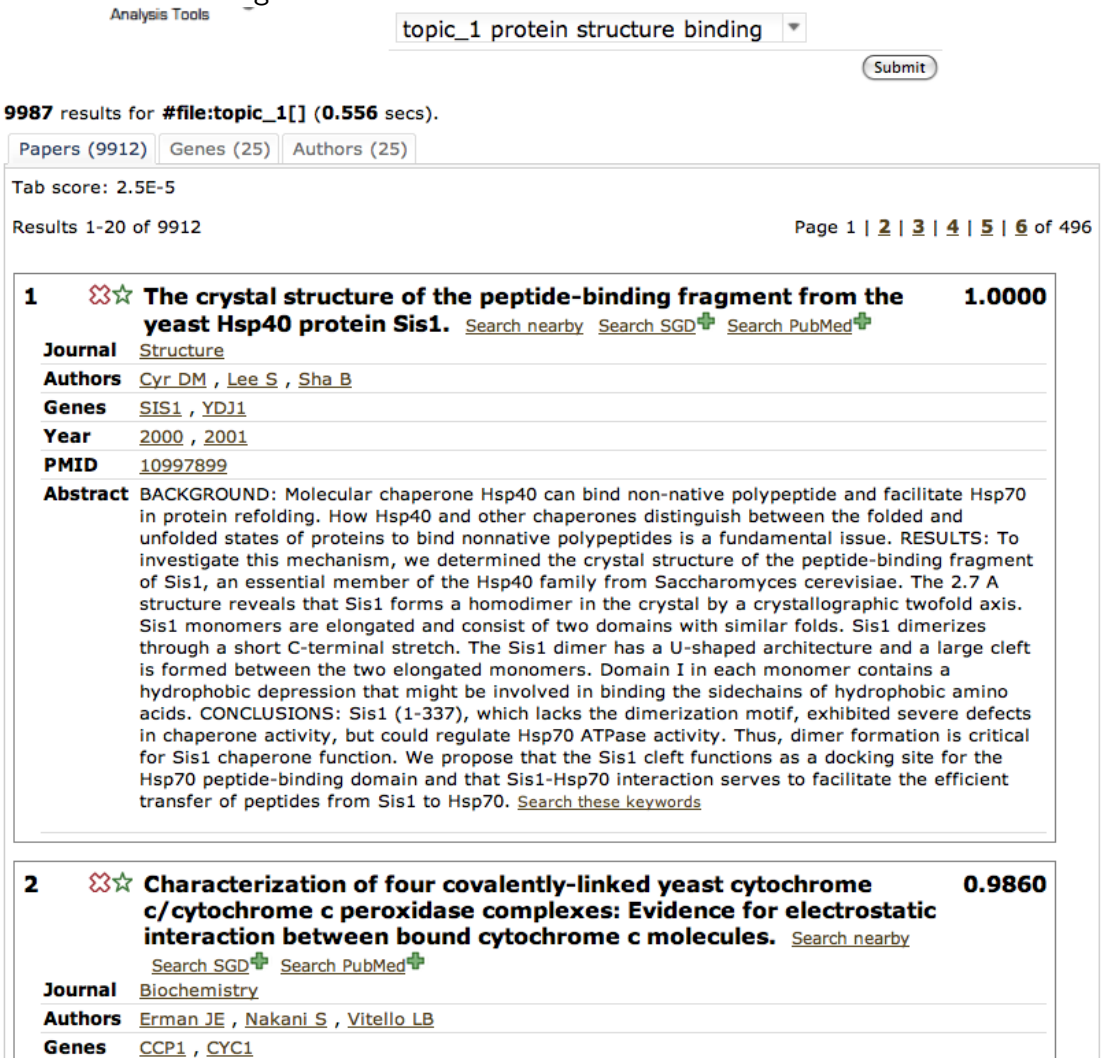
Once the topics are obtained, we first obtain the proteins associated with the topic by retrieving the top proteins from the multinomial distribution corresponding to proteins. Then, the top articles corresponding to each topic are obtained using a ranked list of documents with the highest mass of their topic proportion distributions (θ) residing in the topic being considered.

Manual Evaluation To evaluate the topics, a yeast biologist who is an expert in the field was asked to mark each topic with a binary flag indicating if the top words of the distribution represented a coherent sub-topic in yeast biology. The top words of the distribution representing a topic were presented as a ranked list of words. This process was repeated for the 3 different variants of the model. The variant used to obtain results is concealed from the evaluator to remove the possibility of bias.

In the next step of the evaluation, the top articles and proteins assigned to each topic were presented in a ranked list and a similar judgement was requested to indicate if the article/protein was relevant to the topic in question. Similar to the topic coherence judgements, the process was repeated for each variant of the model. Screenshots of the tool used for obtaining the judgments can be seen in Figure 7.6. It should be noted that since the nature of the topics in the literature considered was highly technical and specialized, it was impractical to get judgements from multiple annotators.

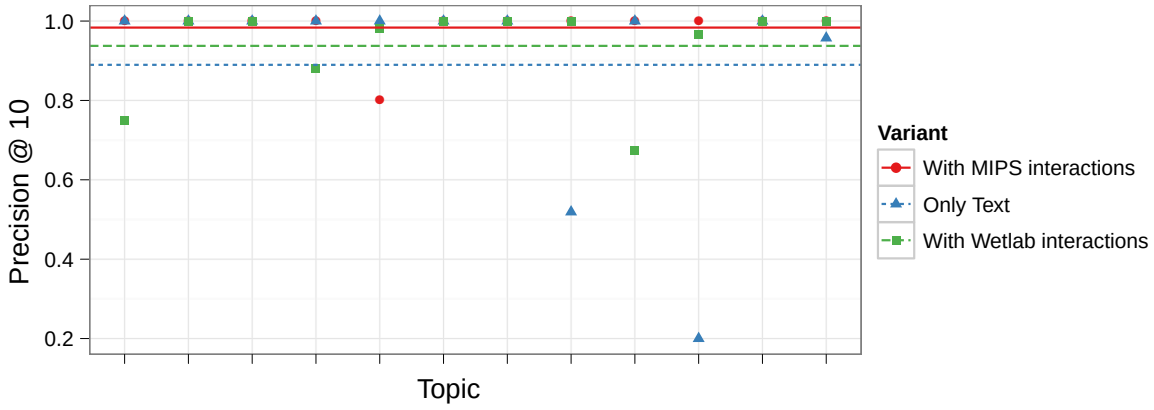
To evaluate the retrieval of the top articles and proteins, we measure the quality of the results by computing its precision@10 score.

Figure 7.6: Screenshot: Article Relevance Annotation Tool

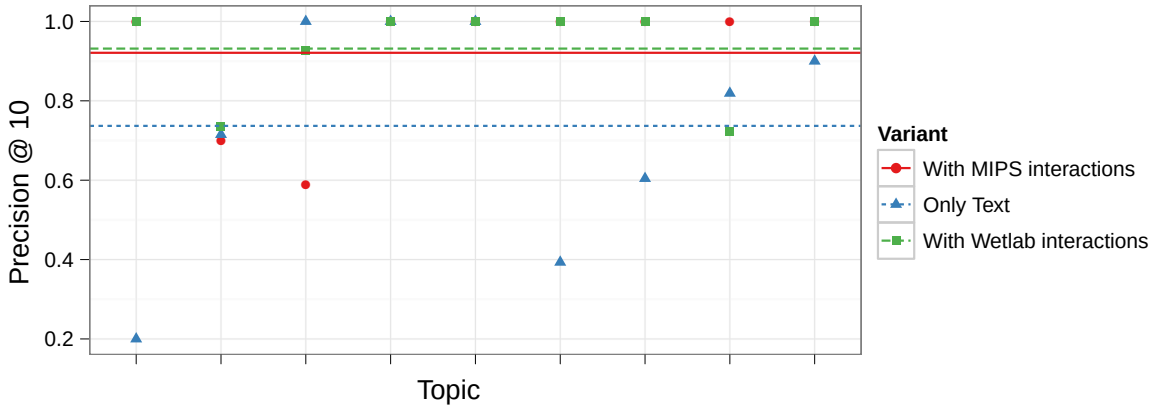


First we evaluate the coherence of the topics obtained from the 3 variants described above. Table 7.3 shows that out of the 15 topics that were obtained, 12 topics were deemed coherent from the text-only model and 13 and 15 topics were deemed coherent from the Block-LDA models using the MIPS and wetlab PPI datasets respectively.

Next, we study the precision@10 values for each topic and variant for the article retrieval and protein retrieval tasks (Figures 7.7(a) and 7.7(b)). The horizontal lines in the plots represent the mean of the precision@10 across all topics. It can be seen from the plots that for both the article and protein retrieval tasks, the joint models work better than the text-only model on average. For the article retrieval task, the model trained with the text + MIPS resulted in the higher mean



(a) Article Retrieval



(b) Protein Retrieval

Figure 7.7: Retrieval Performance

precision@10 whereas for the protein retrieval task, the text + Wetlab PPI dataset returned a higher mean precision@10 value. For both the protein retrieval and paper retrieval tasks, the improvements shown by the joint models using either of the PPI datasets over the text-only model (i.e. the Link LDA model) were statistically significant at the 0.05 level using the paired Wilcoxon sign test. The difference in performance between the two joint models that used the two different PPI networks were, however, insignificant, which indicates that there is no observable advantage in using one PPI dataset over the other in conjunction with the text corpus.

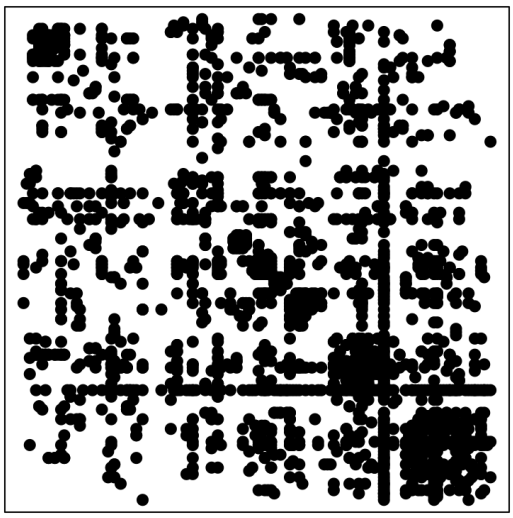
Method	F_1	NMI	Precision	Recall
Regularized Block-LDA	$\mathbf{0.308}^*$	$\mathbf{0.453}^*$	$\mathbf{0.305}^*$	$\mathbf{0.314}^*$
Block-LDA	0.249^*	0.441	0.247^*	0.250^*
Sparse Block model	0.161	0.359^*	0.224^*	0.126
Link LDA	0.152	0.364^*	0.150	0.155^*
MMSB	0.165^*	0.353^*	0.166^*	0.164
Random	0.145	0.241	0.155	0.137

Table 7.4: Functional category prediction (all values computed using 10-fold cross-validation)

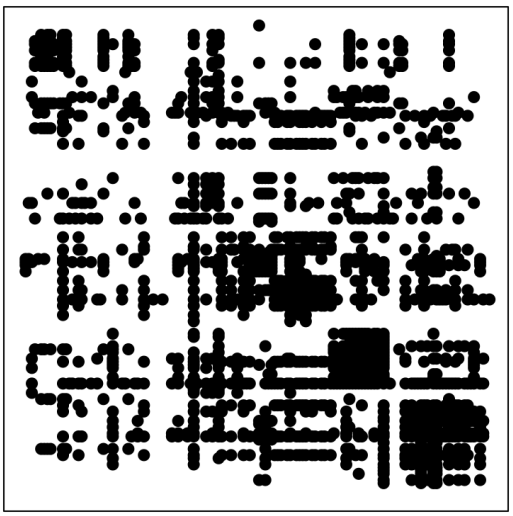
Functional category prediction

Proteins are identified as belonging to multiple functional categories in the MIPS PPI dataset, as described in Section 7.3. We use Block-LDA and baseline methods to predict proteins' functional categories and evaluate it by comparing it to the ground truth in the MIPS dataset using the method presented in prior work [Airoldi et al., 2008]. A model is first trained with K set to 15 topics to recover the 15 top level functional categories of proteins. Every topic that is returned consists of a set of multinomials including $\beta_{\mathbf{t}_1}$, the topic wise distribution over all proteins. The values of $\beta_{\mathbf{t}_1}$ are thresholded such that the top $\approx 16\%$ (the density of the protein-function matrix) of entries are considered as a positive prediction that the protein falls in the functional category corresponding to the latent topic. To determine the mapping of latent topic to functional category, 10% of the proteins are used in a procedure that greedily finds the alignment resulting in the best accuracy, as described in [Airoldi et al., 2008]. It is important to note that the true functional categories of proteins are completely hidden from the model. The functional categories are used only during evaluation of the topics from the model.

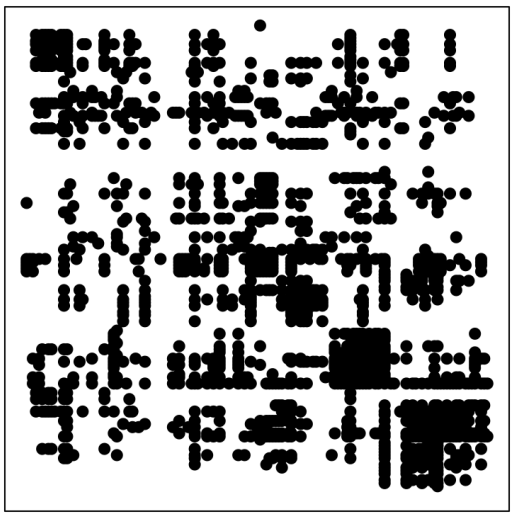
The precision, recall, NMI and F_1 (macro-averaged) scores of the different models in predicting the right functional categories for proteins are shown in Table 7.4. Since there are 15 functional categories and a protein has approximately 2.5 functional category associations, we expect only $\sim 1/6$ of protein-functional category associations to be positive. Precision and recall therefore depict a better picture of the predictions than accuracy. For the random baseline, every protein-functional category pair is randomly deemed to be 0 or 1 with the Bernoulli probability of an



(a) Observed network



(b) From sparse block model



(c) From Block-LDA

Figure 7.8: Enron network and its de-noised recovered versions

association being proportional to the ratio of 1's observed in the protein-functional category matrix in the MIPS dataset. In the MMSB approach, induced latent blocks are aligned to functional categories as described in Airoldi et al. [2008].

We see that the F_1 and NMI scores for the baseline sparse block model and MMSB are nearly the same and that combining text and links provides a significant boost to both scores. Regularization of proteins in text also provides a noticeable improvement. This suggests that protein co-occurrence

patterns in the abstracts contain information about functional categories as is also evidenced by the better than random F_1 score obtained using Link LDA which uses only documents. All the methods considered outperform the random baseline.

7.4.2 Results from the Enron email corpus dataset

As described in Section 7.3, the Enron dataset consists of two components - text from the sent folders and the network of senders and recipients of emails within the Enron organization. Each email is treated as a document and is annotated with a set of people consisting of the senders and recipients of the email. We first study the network reconstruction capability of the Block-LDA model. Block-LDA is trained using all the 96,103 emails in the sent folders and the 200,404 links obtained from the full email corpus. Figures 7.8(a), 7.8(b) and 7.8(c) show the true communication matrix, the matrix reconstructed using the sparse mixed membership stochastic block model and the matrix reconstructed using the Block-LDA model respectively. The figures show that both models are approximately able to recover the communication network in the Enron dataset.

Next, we study the top words and people in the topics induced by Block-LDA shown in Table 7.5. The table shows sample topics induced after running Block-LDA with $K = 15$. The topic labels and notes were hand created after looking at the top words and employees and by using the partial knowledge available about the roles of the employees in the Enron organization [Shetty and Adibi, 2004]. It can be seen that the people within the recovered topics are likely to need to communicate with each other. These instances of topics suggest that the topics capture both notions of semantic concepts obtained from the text of the emails and sets of people who need to interact regularly about the concepts.

Figure 7.9(a) shows the link perplexity and person perplexity in text of held out data, as the number of topics is varied. Person perplexity is indicative of the surprise inherent in observing a sender or a recipient and can be used as a prior in tasks like predicting recipients for emails that are being composed. Link perplexity is a score for the quality of link prediction and captures the notion of social connectivity in the graph. It indicates how well the model is able to capture links between people in the communication network. The person perplexity in the plot decreases initially and stabilizes when the number of topics reaches 20. It eventually starts to rise again when the

Words	contract, party, capacity, gas, df, payment, service, tw, pipeline, issue, rate, section, project, time, system, transwestern, date, el, payment, due, paso
Employees	fossum, scott, harris, hayslett, campbell, geaccone, hyatt, corman, donoho, lokay
Notes	Geaconne was the executive assistant to Hayslett who was the Chief Financial Officer and Treasurer of the Transwestern division of Enron.

(a) Financial contracts

Words	power, california, energy, market, contracts, davis, customers, edison, bill, ferc, price, puc, utilities, electricity, plan, pge, prices, utility, million, jeff
Employees	dasovich, steffes, shapiro, kean, williams, sanders, smith, lewis, wolfe, bass
Notes	Dasovitch was a Government Relations executive, Steffes the VP of government affairs, Shapiro, the VP of regulatory affairs and Haedicke worked for the legal department.

(b) Energy distribution

Words	enron, business, management, risk, team, people, rick, process, time, information, issues, sally, mike, meeting, plan, review, employees, operations, project, trading
Employees	kitchen, beck, lavorato, delainey, buy, presto, shankman, mcconnell, whalley, haedicke
Notes	The people in this topic are top level executives: Kitchen was the President of Enron Online, Beck the Chief operating officer and Lavarato the CEO.

(c) Strategy

Words	deal, deals, dec, mid, book, pst, columbia, please, pl, kate, desk, west, changed, file, questions, mike, report, books, mw, thanks
Employees	love, semperger, symes, giron, keiser, williams, mclaughlin, white, forney, grigsby
Notes	This topic about trading has Semperger in the most likely list of people who was a senior analyst dealing with cash accounts and Forney who worked as a trader at the real time trading desk.

(d) Trading

Words	legal, trading, credit, master, energy, eol, isda, list, counterparty, company, financial, agreement, power, trade, inc, access, products, mark, approval, swap, request
Employees	dasovich, sanders, haedicke, kean, steffes, derrick, harris, williams, shapiro, davis
Notes	As noted before, Dasovich, Haedicke and Steffes performed roles that involved interacting with government agencies.

(e) Legal and regulatory affairs

Words	gas, storage, volumes, volume, demand, capacity, transport, ces, deal, price, day, month, daily, market, ena, contract, power, prices, cash, index
Employees	germany, farmer, grigsby, tholt, townsend, smith, parks, neal, causholli, hernandez
Notes	Farmer was a logistics manager and Tholt was the VP of the division.

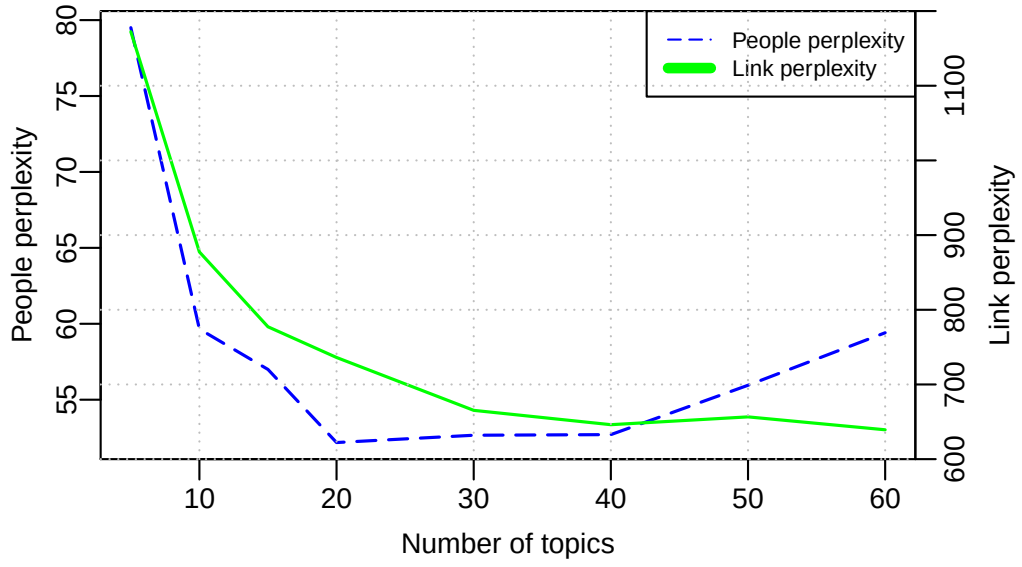
(f) Logistics

Table 7.5: Top words and people from latent topics in the Enron corpus

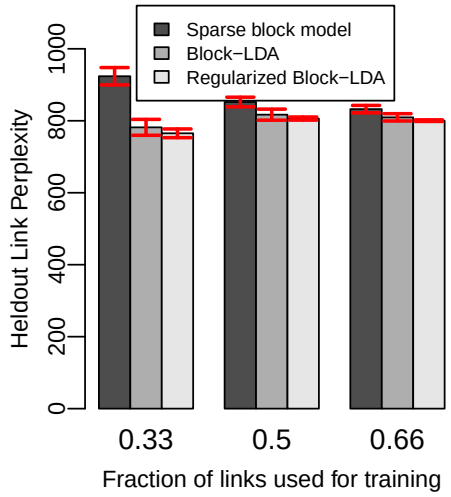
number of topics is raised above 40. The link perplexity on the other hand stabilizes at 20 and then exhibits a slight downward trend. For the remaining experiments with the Enron data, we set $K = 40$.

In the next set of experiments, we evaluate Block-LDA and other models by evaluating the person perplexity in held out emails by varying the training and test set size. Similar to the experiments with the PPI data, the collapsed Gibbs sampler is run until the held out perplexity stabilizes to a nearly constant value ($\approx$ 80 iterations). The perplexity values are averaged over 10 trials. Figure 7.9(c) shows the person perplexity in text in held out data as increasing amounts of the text data are used for training. The remainder of the dataset is used for testing. It is important to note that only Block-LDA uses the communication link matrix. A consistent improvement in person perplexity can be observed when email text data are supplemented with communication link data irrespective of the training set size. This indicates that the latent block structure in the links is beneficial while shaping latent topics from text.

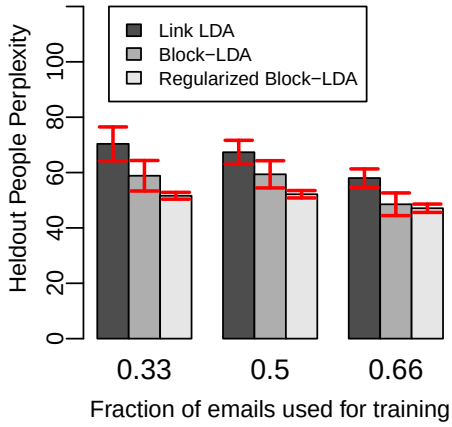
Block-LDA is finally evaluated using link prediction. The sparse block model which serves as a baseline does not use any text information. Figure 7.9(b) shows the perplexity in held out data with varying amounts of the 200,404 edges in the network used for training. When textual information is



(a) Determining number of topics



(b) Heldout link perplexity



(c) Heldout people perplexity

Figure 7.9: Enron corpus: Perplexity (computed using 10-fold cross-validation)

used, all the 96,103 emails are used. The histogram shows that Block-LDA obtains lower perplexities than the sparse block model which uses only links. As in the PPI experiments, using the text in the emails improves the modeling of the network of senders and recipients although the effect is less marked when the number of links used for training is increased. The topical coherence in the

latent topics induces better latent blocks in the matrix, indicating a transfer of signal from the text to the network model.

7.5 Related work

Link LDA and many other extensions to LDA model documents that are annotated with metadata. In a parallel area of research, various different approaches to modeling links between documents have been explored. For instance, Pairwise-Link-LDA [Nallapati et al., 2008] combines MMSB with LDA by modeling documents using LDA and generating links between them using MMSB. The Relational Topic Model [Chang and Blei, 2009b] generates links between documents based on their topic distributions. The Copycat and Citation Influence models [Dietz et al., 2007] also model links between citing and cited documents by extending LDA and eliminating independence between documents. The Latent Topic Hypertext Model (LTHM) [Gruber et al., 2008] presents a generative process for documents that can be linked to each other from specific words in the citing document. The model proposed here, Block-LDA, is different from this class of models in that they model links between entities in the documents rather than links between documents.

The Nubbi model [Chang et al., 2009] tackles a related problem where entity relations are discovered from text data by relying on words that appear in the context of entities and entity pairs in the text. Block-LDA differs from Nubbi in that it models a document as bags of entities without considering the location of entity mentions in the text. The entities need not even be mentioned in the text of the document. The Group-Topic model [Wang et al., 2006] addresses the task of modeling events pertaining to pairs of entities with textual attributes that annotate the event. The text in this model is associated with events, which differs from the standalone documents mentioning entities considered by Block-LDA.

The Author-Topic model (AT) [Rosen-Zvi et al., 2004] addresses the task of modeling corpora annotated with the ids of people who authored the documents. Every author in the corpus has a topic distribution over the latent topics, and words in the documents are drawn from topics drawn from the specific distribution of the author who is deemed to have generated the word. The Author-Recipient-Topic model [McCallum et al., 2005] extends the idea further by building a topic

Model	Links	Documents
LDA [Blei et al., 2003]	-	words
Link LDA [Erosheva et al., 2004]	-	words + entities
Relational Topic model [Chang and Blei, 2009a]	document-document	words + document ids
Pairwise Link-LDA [Nallapati et al., 2008], Link-PLSA-LDA [Nallapati and Cohen, 2008]	document-document	words + cited document ids
Copycat, Citation Influence models [Dietz et al., 2007]	document-document	words + cited document ids
Latent Topic Hypertext model [Gruber et al., 2008]	document-document	words + cited document ids
Author Recipient Topic model [McCallum et al., 2005]	-	docs + authors + recipients
Author Topic model [Rosen-Zvi et al., 2004]	-	docs + authors
Topic Link LDA [Liu and Niculescu-mizil, 2009]	document-document	words + authors
MMSB [Airoldi et al., 2008]	entity-entity	-
Sparse block model [Parkkinen et al., 2009]	entity-entity	-
Nubbi [Chang et al., 2009]	entity-entity	words near entities or entity-pairs
Group topic model [Wang et al., 2006]	entity-entity	words about the entity-entity event
Block-LDA [Balasubramanyan and Cohen, 2011]	entity-entity	words + entities

Table 7.6: Related work

distribution for every author-recipient pair. As we show in the experiments below, Block-LDA can also be used to model the relationships between authors, recipients, and words in documents, by

constructing an appropriate link matrix from known information about the authors and recipients of documents; however, unlike the AT and ART models, which are primarily designed to model documents, Block-LDA provides a generative model for the links between authors and recipients in addition to documents. This allows Block-LDA to be used for additional inferences not possible with the AT or ART models, for instance, predicting probable author-recipient interactions. Wen and Lin [2010] describe an application of an approach that uses both content and network information to analyse enterprise data. While a joint modeling of the network and content is not used, LDA is used to study the topics in communications between people. More recently, Zhang and Carin [2012] presented an approach that jointly models matrices and related documents using a model based on the Indian Buffet Process. This model shares many similarities with Block-LDA in that topics represent both blocks in the matrix and a distribution over words in documents. It however differs from Block-LDA in that the matrix in their approach represents external data about documents unlike our scenario, where the matrix represents relations between entities tagged in the documents. Moreover, they use focused topic models [Williamson et al., 2010] and binary matrix factorization [Meeds et al., 2007] to model the document corpus ahd matrix respectively instead of Link-LDA and the PSK model as in Block-LDA.

A summary of related models from prior work is shown in Table 7.6.

7.6 Conclusion

The work in this chapter is an extension of work presented earlier in Balasubramanyan and Cohen [2011]. We proposed a model that performs data fusion by jointly models links between entities and text annotated with entities that permits co-occurrence information in text to influence link modeling and vice versa. Our experiments show that joint modeling outperforms approaches that use only a single source of information. Improvements are observed when the joint model is evaluated internally using perplexity in two different datasets and externally using protein functional category prediction in the yeast dataset. Moreover, the topics induced by the model when examined subjectively appear to be useful in understanding the structure of the data both in terms of the topics discussed and in terms of the connectivity characteristics between entities.

Chapter 8

Conclusion

8.1 Conclusion

Probabilistic models for text modeling and network modeling are widely used. Mixed-membership models especially, have become quite popular within the last decade. While, mixed-membership models are convenient to capture complexity, it is not always obvious that the full power of mixed-membership is always necessary. The main motivation in this work is in controlling the degree of mixed-membership. To this end, we presented techniques to exercise fine-level control over the latent structure uncovered in mixed-membership models. We explored the fully Bayesian approach using the ECD prior and determined that it can be approximated well using a regularizer that offers computational advantages. The regularization framework that we presented, enabled placing of preferences on functions over aggregate values of latent variable assignments with only limited computational cost. The framework's utility was demonstrated in both text models (i.e. topic models) and network models. In topic models, we used the entropic regularization approach to obtain slightly mixed-membership models in which documents and words have limited ability to span different latent roles. Next, we also presented a method to incorporate limited labeled data into mixed-membership models by modifying the collapsed Gibbs sampling approximate inference procedure.The utility of the regularization and semi-supervised models was demonstrated in the task of clustering entities obtained from HTML tables.

The regularization approach was next applied to stochastic block models. The utility of the

regularization was evaluted by evaluating the ability of the models to perform cluster recovery. We also presented a model that jointly models networks and text. The joint model was applied to yeast protein-protein networks and literature and evaluated using domain experts.

8.2 Future Work

In this thesis we have focused on using parametric mixed membership models. The regularization framework introduced here can also be applied to non-parametric models [Teh et al., 2006] which remove the need to *apriori* set the number of topics or clusters. The interplay between the concentration parameter in non-parametric models, that control the freedom to create new topics and the regularization will be interesting to consider.

Collapsed Gibbs sampling is the approximate inference technique that we have employed for LDA-like models. A major drawback to collapsed Gibbs sampling is the difficulty of determining whether the sampler has "mixed", i.e. , has reached a stationary distribution. Typically, this is overcome by empirically observing that the sampler has stabilized after a few hundred or thousand iterations. In future work, we intend to study the convergence properties of the collapsed Gibbs sampling technique for topic models with both the regularized and unregularized versions.

In the regularization framework introduced in this thesis, the variance hyperparameter $\sigma_{l_w}^2$ is set to a fixed value. In future work, we wish to use an empirical Bayes inference procedure to optimize the value of this parameter. The procedure will effectively alternate between sampling values for the z variables and then estimating values for the hyperparameter in an EM-like approach. We also wish to investigate the interaction between the Dirichlet and variance hyperparameters in detail. Our experiments (figure 3.8) with the reviews datasets indicated that regularization provides a better performance gain in MSE than tuning the Dirichlet hyperparameter. The generalizability of this gain requires further investigation.

The third line of work we wish to pursue in the future is in the realm of semi-supervised learning. Our experiments demonstrated that using even a few labeled documents and features provides noticeable improvements. In future work, we wish to get a detailed comparison on the benefits of using labeled documents versus labeled features. If we had a set budget in terms for

resources for labeling, we would like to produce a framework which indicates what the best use of the resources is, *vis-a-vis* labeling documents or features. This line of work is especially interesting considering the increasingly wide use of Amazon's Mechanical Turk for annotations purposes.

Finally, we wish to investigate the issues of scale inherent in the class of mixed-membership models. Scalability is critical as we start to apply mixed-membership modeling to large knowledge databases like NELL [Carlson et al., 2010b]. Parallelization of approximate inference [Newman et al., 2006b] has been an active area of research and we wish to build on the advances on that front, to scale inference using topic models that are regularized using the framework proposed here.

Bibliography

Lada A. Adamic and Natalie Glance. The political blogosphere and the 2004 u.s. election: divided they blog. In *LinkKDD '05: Proceedings of the 3rd international workshop on Link discovery*, pages 36–43, New York, NY, USA, 2005. ACM. ISBN 1-59593-215-1. doi: http://doi.acm.org/10.1145/1134271.1134277. 6.5.1

Edoardo M. Airoldi, David Blei, Stephen E. Fienberg, and Eric P. Xing. Mixed membership stochastic blockmodels. *Journal of Machine Learning Research*, 9:1981–2014, September 2008. 1.1, 6.1, 6.2, 6.6, 7.1, 7.2, 7.4.1, 7.5

David Andrzejewski and David Buttler. Latent topic feedback for information retrieval. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '11, pages 600–608, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0813-7. doi: 10.1145/2020408.2020503. URL http://doi.acm.org/10.1145/2020408.2020503. 1.1

David Andrzejewski and Xiaojin Zhu. Latent Dirichlet Allocation with topic-in-set knowledge. pages 43–48, June 2009. URL http://dl.acm.org/citation.cfm?id=1621829.1621835. 5.6

David Andrzejewski, Xiaojin Zhu, and Mark Craven. Incorporating domain knowledge into topic modeling via Dirichlet Forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pages 1–8, New York, New York, USA, June 2009. ACM Press. ISBN 9781605585161. doi: 10.1145/1553374.1553378. URL http://dl.acm.org/citation.cfm?id=1553374.1553378. 5.6

Rachit Arora and Balaraman Ravindran. Latent Dirichlet Allocation and singular value decomposition based multi-document summarization. In *ICDM*, pages 713–718. IEEE Computer Society, 2008. 1.1

Arthur Asuncion, Max Welling, Padhraic Smyth, and Yee Whye Teh. On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 27–34, Arlington, Virginia, United States, 2009. AUAI Press. ISBN 978-0-9749039-5-8. URL http://dl.acm.org/citation.cfm?id=1795114.1795118. 2.1

Josh Attenberg, Prem Melville, and Foster Provost. A unified approach to active dual supervision for labeling features and examples. In *Proceedings of the 2010 European conference on Machine learning and knowledge discovery in databases: Part I*, ECML PKDD'10, pages 40–55, Berlin, Heidelberg, 2010. Springer-Verlag. ISBN 3-642-15879-X, 978-3-642-15879-7. 5.6

Anton Bakalov, Andrew McCallum, Hanna Wallach, and David Mimno. Topic models for taxonomies. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries - JCDL '12*, page 237, New York, New York, USA, June 2012. ACM Press. ISBN 9781450311540. doi: 10.1145/2232817.2232861. URL http://dl.acm.org/citation.cfm?id=2232817.2232861. 5.6

Ramnath Balasubramanyan and William W. Cohen. Block-LDA: Jointly modeling entity-annotated text and entity-entity links. In *SDM*, pages 450–461. SIAM / Omnipress, 2011. ISBN 978-0-898719-92-5. 6.1, 6.2, 7.1, 7.5, 7.6

Ramnath Balasubramanyan and William W. Cohen. Regularization of latent variable models to obtain sparsity. In *SDM*, 2013. 3.7

Ramnath Balasubramanyan, Frank Lin, and William W. Cohen. Node clustering in graphs: An empirical study. In *NIPS Workshop on Networks Across Disciplines in Theory and Applications*, 2010. 1.1, 6.5.1, 6.7

Ramnath Balasubramanyan, Bhavana Dalvi, and William W. Cohen. From topic models to semi-supervised learning: Biasing mixed-membership models to exploit topic-indicative features in entity clustering. In *Proceedings of the 2013 European conference on Machine Learning and Knowledge Discovery in Databases*, ECML PKDD'13. Springer-Verlag, 2013. 5.2, 5.7

David Blei and Jon McAuliffe. *Supervised Topic Models*, pages 121–128. MIT Press, Cambridge, MA, 2008. 2, 2.2, 5.6

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet Allocation. *The Journal of*

Machine Learning Research, 3:993–1022, 2003. 1.1, 2, 2.1, 2.1, 3.1, 6.1, 7.1, 7.5

John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, Bollywood, boomboxes and blenders: Domain adaptation for sentiment classification. In *ACL*, pages 187–205, 2007. 3.4

Deng Cai, Qiaozhu Mei, Jiawei Han, and Chengxiang Zhai. Modeling hidden topics on document manifold. In *Proceeding of the 17th ACM conference on Information and knowledge mining - CIKM '08*, page 911, New York, New York, USA, October 2008. ACM Press. 3.1, 3.6

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr., and Tom M. Mitchell. Toward an architecture for never-ending language learning. In *Proceedings of the Twenty-Fourth Conference on Artificial Intelligence (AAAI 2010)*, 2010a. 5.2

Andrew Carlson, Justin Betteridge, Richard C. Wang, Estevam R. Hruschka, Jr., and Tom M. Mitchell. Coupled semi-supervised learning for information extraction. In *Proceedings of the third ACM international conference on Web search and data mining*, WSDM '10, pages 101–110, New York, NY, USA, 2010b. ACM. ISBN 978-1-60558-889-6. doi: 10.1145/1718487.1718501. URL http://doi.acm.org/10.1145/1718487.1718501. 5.2, 8.2

Gilles Celeux and Gilda Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, 13(2):195–212, 1996. 3.6

J. Chang and D. M Blei. Relational topic models for document networks. 2009a. URL https://www.cs.princeton.edu/~blei/papers/ChangBlei2009.pdf. 7.5

Jonathan Chang. Gibbs sampling equations for LDA and related models. Derivation, 2011. URL http://lists.cs.princeton.edu/pipermail/topic-models/attachments/20110210/89b1646c/attachment-0001.pdf. 2.2

Jonathan Chang and David M. Blei. Relational topic models for document networks. In *Proc. of Conf. on AI and Statistics (AISTATS 09)*, 2009b. 7.5

Jonathan Chang and David M. Blei. Hierarchical relational models for document networks. *Annals of Applied Statistics*, October 2010. URL http://arxiv.org/abs/0909.4331. 3.6

Jonathan Chang, Jordan Boyd-Graber, and David M. Blei. Connections between the lines: augmenting social networks with text. In *Proceedings of the 15th ACM SIGKDD international*

conference on Knowledge discovery and data mining, page 169178, 2009. 7.5

Ming-Wei Chang, Lev Ratinov, and Dan Roth. Guiding semi-supervision with constraint-driven learning. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 280–287, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P07-1036. 3.6

Adrian Corduneanu and Tommi S Jaakkola. Distributed information regularization on graphs. *Advances in Neural Information Processing Systems 17*, .17:297–304, 2005. 3.6

Bhavana Bharat Dalvi, William W. Cohen, and Jamie Callan. Websets: extracting sets of entities from the web using unsupervised information extraction. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 243–252, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-0747-5. doi: 10.1145/2124295.2124327. URL http://doi.acm.org/10.1145/2124295.2124327. 5.1, 5.2, 5.2, 5.5, 5.6

Laura Dietz, Steffen Bickel, and Tobias Scheffer. Unsupervised prediction of citation influences. *Proceedings of the 24th Annual International Conference on Machine Learning (ICML 2007)*, page 233240, 2007. 7.5

S.S. Dwight, R. Balakrishnan, K.R. Christie, M.C. Costanzo, K. Dolinski, S.R. Engel, B. Feierbach, D.G. Fisk, J. Hirschman, E.L. Hong, et al. Saccharomyces genome database: Underlying principles and organisation. *Briefings in bioinformatics*, 5(1):9, 2004. 7.3

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. Sparse additive generative models of text. In Lise Getoor and Tobias Scheffer, editors, *ICML*, pages 1041–1048. Omnipress, 2011. 3.6

Elena A. Erosheva, Stephen Fienberg, and John Lafferty. Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5220, 2004. 1.1, 2, 2.1, 5.2, 7.1, 7.5

Kuzman Ganchev, João Graça, Jennifer Gillenwater, and Ben Taskar. Posterior regularization for structured latent variable models. *J. Mach. Learn. Res.*, 11:2001–2049, August 2010. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1756006.1859918. 3.6

Zoubin Ghahramani and Matthew J Beal. Graphical models and variational methods. *Advanced Mean Field MethodTheory and Practice*, 2000. 2.1

Anna Goldenberg, Alice X. Zheng, Stephen E. Fienberg, and Edoardo M. Airoldi. A survey of statistical network models. *Found. Trends Mach. Learn.*, 2:129–233, February 2010. ISSN 1935-8237. 6.1

Yves Grandvalet and Yoshua Bengio. Semi-supervised Learning by Entropy Minimization. *Advances in neural information processing systems*, 17:529–536, 2005. 3.6

Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proc Natl Acad Sci U S A*, 101 Suppl 1:5228–5235, April 2004. ISSN 0027-8424. 1.1, 2.1, 2.1.1, 7.1

Thomas L. Griffiths, Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. Integrating topics and syntax. In *In Advances in Neural Information Processing Systems 17*, pages 537–544. MIT Press, 2005. 1.1

Amit Gruber, Michal Rosen-zvi, and Yair Weiss. Latent topic models for hypertext. *UAI*, 2008. URL http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.153.6378http://www.cs.huji.ac.il/~amitg/uai08.pdf. 3.6, 7.5

Marti A. Hearst. Automatic acquisition of hyponyms from large text corpora. In *Proceedings of the 14th conference on Computational linguistics - Volume 2*, COLING '92, pages 539–545, Stroudsburg, PA, USA, 1992. Association for Computational Linguistics. doi: 10.3115/992133.992154. URL http://dx.doi.org/10.3115/992133.992154. 5.2

Gregor Heinrich. Parameter estimation for text analysis. *Bernoulli*, 2009. 2.1.1

Qirong Ho, Junming Yin, and Eric P. Xing. On triangular versus edge representations — towards scalable modeling of networks. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger, editors, *NIPS*, pages 2141–2149, 2012. 6.6

Paul W. Holland, Kathryn B. Laskey, and Samuel Leinhardt. Stochastic blockmodels: First steps. *Social Networks*, 5(2):109–137, 1983. ISSN 03788733. 1.1, 6.1

Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177, New York, NY, USA, 2004. ACM. ISBN 1-58113-888-1. doi: 10.1145/1014052.1014073. URL http://doi.acm.org/10.1145/1014052.1014073. 3.6

Sonia Jain and Radford Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13:158–182, 2000. 4.4

Feng Jiao, Shaojun Wang, Chi-Hoon Lee, Russell Greiner, and Dale Schuurmans. Semi-supervised conditional random fields for improved sequence segmentation and labeling. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the ACL - ACL '06*, pages 209–216, Morristown, NJ, USA, July 2006. Association for Computational Linguistics. 3.6

Yohan Jo and Alice H. Oh. Aspect and sentiment unification model for online review analysis. In *Proceedings of the fourth ACM International Conference on Web search and Data Mining - WSDM '11*, page 815, New York, New York, USA, February 2011. ACM Press. ISBN 9781450304931. doi: 10.1145/1935826.1935932. URL http://dl.acm.org/citation.cfm?id=1935826.1935932. 3.4

Mahesh Joshi, Mark Dredze, William W. Cohen, and Carolyn Penstein Rosé. Multi-domain learning: When do domains matter? In *EMNLP-CoNLL*, pages 1302–1312. ACL, 2012. ISBN 978-1-937284-43-5. 3.4

Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97, 1955. ISSN 1931-9193. doi: 10.1002/nav.3800020109. 6.5.1, 7.4.1

Kevin Lerman, Sasha Blair-Goldensohn, and Ryan McDonald. Sentiment summarization: evaluating and learning user preferences. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 514–522, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1609067.1609124. 3.6

Jure Leskovec, Deepayan Chakrabarti, Jon Kleinberg, Christos Faloutsos, and Zoubin Ghahramani. Kronecker graphs: An approach to modeling networks. *J. Mach. Learn. Res.*, 11:985–1042, March 2010a. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1756006.1756039. 6.6

Jure Leskovec, Kevin J. Lang, and Michael Mahoney. Empirical comparison of algorithms for

network community detection. In *Proceedings of the 19th international conference on World wide web - WWW '10*, page 631, New York, New York, USA, April 2010b. ACM Press. ISBN 9781605587998. 1.1

Percy Liang, Michael I. Jordan, and Dan Klein. Learning from measurements in exponential families. In *Proceedings of the 26th Annual International Conference on Machine Learning - ICML '09*, pages 1–8, New York, New York, USA, June 2009. ACM Press. ISBN 9781605585161. doi: 10.1145/1553374.1553457. URL http://dl.acm.org/citation.cfm?id=1553374.1553457. 3.6

Yan Liu and Alexandru Niculescu-mizil. Topic-Link LDA : Joint Models of Topic and Author Community. *Work*, 2009. 7.5

Qing Lu and Lise Getoor. Link-based classification. In Tom Fawcett and Nina Mishra, editors, *ICML*, pages 496–503. AAAI Press, 2003. ISBN 1-57735-189-4. 6.5.1

Ulrike Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, August 2007. ISSN 0960-3174. 1.1, 6.1

David J. C. MacKay. *Information Theory, Inference & Learning Algorithms*. Cambridge University Press, New York, NY, USA, 2002. ISBN 0521642981. 5.2

Gideon S Mann and Andrew McCallum. Generalized Expectation Criteria for semi-supervised learning with weakly labeled data. *Journal of Machine Learning Research*, 11:955–984, 2010. ISSN 15324435. URL http://dl.acm.org/citation.cfm?id=1756038. 3.6

Xian-Ling Mao, Zhao-Yan Ming, Tat-Seng Chua, Si Li, Hongfei Yan, and Xiaoming Li. SSHLDA: a semi-supervised hierarchical topic model. pages 800–809, July 2012. URL http://dl.acm.org/citation.cfm?id=2390948.2391034. 5.6

Lluis Marquez, Gerard Escudero, David Martinez, and German Rigau. WSD in NLP Applications. In Eneko Agirre and Phil Edmonds, editors, *Word Sense Disambiguation Algorithms and Applications*, volume 33 of *Text, Speech and Language Technology*, chapter 7, pages 167–208. Springer, 2006. 3.1

Andrew McCallum, Andres Corrada-Emmanuel, and Xuerui Wang. Topic and role discovery in social networks. *IN IJCAI*, pages 786—791, 2005. URL http://citeseerx.ist.psu.edu/

viewdoc/summary?doi=10.1.1.110.7687. 7.5, 7.5

Edward Meeds, Zoubin Ghahramani, Radford M. Neal, and Sam T. Roweis. Modeling dyadic data with binary latent factors. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 977–984. MIT Press, Cambridge, MA, 2007. 7.5

Qiaozhu Mei, Deng Cai, Duo Zhang, and ChengXiang Zhai. Topic modeling with network regularization. In *Proceeding of the 17th international conference on World Wide Web - WWW '08*, page 101, New York, New York, USA, April 2008. ACM Press. 3.6

Hans-Werner Mewes, C. Amid, Roland Arnold, Dmitrij Frishman, Ulrich Gldener, Gertrud Mannhaupt, Martin Mnsterktter, Philipp Pagel, Normann Strack, Volker Stmpflen, Jens Warfsmann, and Andreas Ruepp. Mips: Analysis and annotation of proteins from whole genomes. *Nucleic Acids Res*, 32:41–44, 2004. 7.3

David M. Mimno and Andrew McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In David A. McAllester and Petri Myllymäki, editors, *UAI*, pages 411–418. AUAI Press, 2008. ISBN 0-9749039-4-9. 5.6

Saeedeh Momtazi and Dietrich Klakow. A word clustering approach for language model-based sentence retrieval in question answering systems. In *Proceedings of the 18th ACM conference on Information and knowledge management*, CIKM '09, pages 1911–1914, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-512-3. doi: 10.1145/1645953.1646263. URL http://doi.acm.org/10.1145/1645953.1646263. 3.1

R Nallapati and W Cohen. Link-PLSA-LDA: A new unsupervised model for topics and influence of blogs. Association for the Advancement of Artificial Intelligence, 2008. 7.5

Ramesh M. Nallapati, Amr Ahmed, Eric P. Xing, and William W. Cohen. Joint latent topic models for text and citations. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 542–550, Las Vegas, Nevada, USA, 2008. ACM. 2, 2.1, 7.5

David Newman, Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers. Statistical entity-topic models. In *In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 680–686. ACM Press, 2006a. 2.1.1

David Newman, Padhraic Smyth, and Mark Steyvers. Scalable parallel topic models. *Journal of Intelligence Community Research and Development*, 5, 2006b. 8.2

David Newman, Edwin V. Bonilla, and Wray L. Buntine. Improving topic coherence with regularized topic models. In *NIPS*, pages 496–504, 2011. 3.6

Hwee Tou Ng and Hian Beng Lee. Dso corpus of sense-tagged english, 1997. 3.1

Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Mach. Learn.*, 39(2-3):103–134, May 2000. ISSN 0885-6125. URL http://dx.doi.org/10.1023/A:1007692713085. 5.6

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*, pages 115–124, 2005. 3.4, 3.6

Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2): 1–135, January 2008. ISSN 1554-0669. doi: 10.1561/1500000011. URL http://dx.doi.org/10.1561/1500000011. 3.6

Juuso Parkkinen, Janne Sinkkonen, Adam Gyenge, and Samuel Kaski. A block model suitable for sparse graphs. *The 7th International Workshop on Mining and Learning with Graphs*, 2009. 1.1, 6.1, 6.2, 7.1, 7.2, 7.5

Marius Pasca and Benjamin Van Durme. Weakly-Supervised Acquisition of Open-Domain Classes and Class Attributes from Web Documents and Query Logs. In *Proceedings of the 46th annual meeting of the ACL - ACL '08*, pages 19–27, Columbus, Ohio, June 2008. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P/P08/P08-1003. 5.6

Ana-Maria Popescu and Oren Etzioni. Extracting product features and opinions from reviews. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 339–346, Stroudsburg, PA, USA, 2005. Association for Computational Linguistics. doi: 10.3115/1220575.1220618. URL http://dx.doi.org/10.3115/1220575.1220618. 3.6

Ian Porteous, David Newman, Alexander Ihler, Arthur Asuncion, Padhraic Smyth, and Max Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 569–577,

2008. 2.1

Lizhen Qu, Georgiana Ifrim, and Gerhard Weikum. The bag-of-opinions method for review rating prediction from sparse text patterns. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 913–921, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. URL http://dl.acm.org/citation.cfm?id=1873781.1873884. 3.4, 3.6

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, page 248256, Singapore, August 2009. Association for Computational Linguistics. 5.5, 5.6, 5.6

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 494, 487. AUAI Press, 2004. ISBN 0974903906. URL http://portal.acm.org/citation.cfm?id=1036843.1036902. 7.5, 7.5

TimothyN. Rubin, America Chambers, Padhraic Smyth, and Mark Steyvers. Statistical topic models for multi-label document classification. *Machine Learning*, 88(1-2):157–208, 2012. ISSN 0885-6125. doi: 10.1007/s10994-011-5272-5. URL http://dx.doi.org/10.1007/s10994-011-5272-5. 3.1, 5.6

Gerard Salton and Michael J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., New York, NY, USA, 1986. ISBN 0070544840. 2.1

Burr Settles. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D11-1136. 5.6

Jitesh Shetty and Jafar Adibi. The enron email dataset database schema and brief statistical report, 2004. 7.3, 7.4.2

Jianbo Shi and Jitendra Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000. ISSN 01628828. 1.1, 6.1, 6.4

Tom AB Snijders. Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2):1–40, 2002. 6.6

Tom A.B. Snijders and Krzysztof Nowicki. Estimation and prediction for stochastic blockmodels for graphs with latent block structure. *Journal of Classification*, 14(1):75–100, January 1997. ISSN 0176-4268. 1.1, 6.1

Harold Steck and Tommi S. Jaakkola. On the Dirichlet prior and Bayesian regularization. In *In Advances in Neural Information Processing Systems 15*, pages 697–704. MIT Press, 2002. 4.4

Mark Steyvers, Padhraic Smyth, and Chaitanya Chemuduganta. Combining background knowledge and learned topics. *Topics in Cognitive Science*, 3(1):18–47, 2011. ISSN 17568757. doi: 10.1111/j.1756-8765.2010.01097.x. URL http://doi.wiley.com/10.1111/j.1756-8765.2010.01097.x. 5.6

Partha Pratim Talukdar, Joseph Reisinger, Marius Pasca, Deepak Ravichandran, Rahul Bhagat, and Fernando Pereira. Weakly-Supervised acquisition of labeled class instances using graph random walks. In *Proceedings of the 46th annual meeting of the ACL - ACL '08*, pages 582–590, Honolulu, Hawaii, October 2008. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D08-1061. 5.6

Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006. 8.2

Ivan Titov and Ryan McDonald. A Joint Model of Text and Aspect Ratings for Sentiment Summarization. In *Proceedings of the 46th annual meeting of the ACL - ACL '08*, pages 308–316, Columbus, Ohio, June 2008a. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/P/P08/P08-1036. 3.4, 3.6

Ivan Titov and Ryan McDonald. Modeling online reviews with multi-grain topic models. In *Proceedings of the 17th International Conference on World Wide Web (WWW)*, pages 112–120, Beijing, China, 2008b. 3.4

Chong Wang and D Blei. Variational inference in nonconjugate models. *Journal of Machine Learning Research*, 14:899–925, 2013. 4.4

Chong Wang and David M. Blei. Decoupling sparsity and smoothness in the discrete hierarchical

dirichlet process. In Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta, editors, *NIPS*, pages 1982–1989. Curran Associates, Inc., 2009. ISBN 9781615679119. 3.6

Richard C. Wang and William W. Cohen. Automatic set instance extraction using the web. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 441–449, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics. ISBN 978-1-932432-45-9. URL http://dl.acm.org/citation.cfm?id=1687878.1687941. 5.2

Xuerui Wang, Natasha Mohanty, and Andrew McCallum. Group and topic discovery from relations and their attributes. *In Advances in Neural Information Processing Systems 18*, pages 1449—1456, 2006. 7.5

Yang Wang and Greg Mori. Human action recognition by semilatent topic models. *IEEE transactions on pattern analysis and machine intelligence*, 31(10):1762–74, October 2009. ISSN 0162-8828. doi: 10.1109/TPAMI.2009.43. URL http://ieeexplore.ieee.org/articleDetails.jsp?arnumber=4785474. 5.6

Yang Wang, Payam Sabzmeydani, and Greg Mori. Semi-latent Dirichlet allocation: A hierarchical model for human action recognition. pages 240–254, October 2007. URL http://dl.acm.org/citation.cfm?id=1785357.1785377. 5.6

Zhen Wen and Ching-Yung Lin. Towards finding valuable topics. In *SDM*, pages 720–731, 2010. 7.5

Robert Wetzker, Carsten Zimmermann, and Christian Bauckhage. Analyzing social bookmarking systems: A del.icio.us cookbook. In *Mining Social Data (MSoDa) Workshop Proceedings*, pages 26–30. ECAI 2008, July 2008. URL http://robertwetzker.com/wp-content/uploads/2008/06/wetzker_delicious_ecai2008_final.pdf. 5.2

Sinead Williamson, Chong Wang, Katherine A. Heller, and David M. Blei. The IBP compound Dirichlet process and its application to focused topic modeling. In Johannes Fürnkranz and Thorsten Joachims, editors, *Proceedings of the 27th International Conference on Ma-*

chine Learning (ICML-10), pages 1151–1158, Haifa, Israel, June 2010. Omnipress. URL http://www.icml2010.org/papers/397.pdf. 7.5

Reza Zafarani and Huan Liu. Social computing data repository at ASU, 2009. 6.5.1

XianXian Zhang and Lawrence Carin. Joint Modeling of a Matrix with Associated Text via Latent Binary Features. In *Advances in Neural Information Processing Systems 25*, 2012. URL http://nips.cc/Conferences/2012/Program/event.php?ID=3394. 7.5

Jun Zhu, Amr Ahmed, and Eric P. Xing. MedLDA: maximum margin supervised topic models. *The Journal of Machine Learning Research*, 13(1):2237–2278, January 2012. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2503308.2503315. 5.6